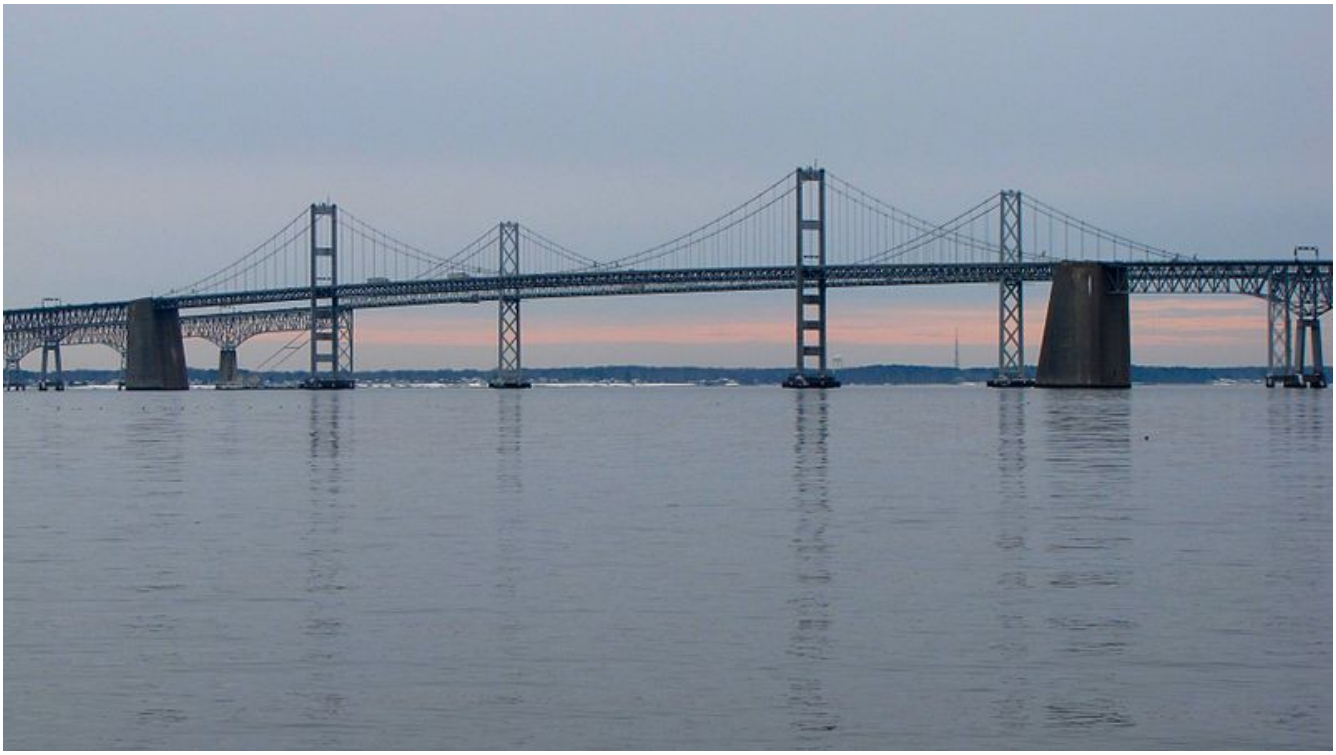
Chesapeake Bay Water Quality

Stacie Kuamoo

CS171 – Spring 2015



Credit: Ben Schumin

Overview and Motivation

I initially chose to do a project visualizing the decline of the oyster population in the Chesapeake Bay in recent years. I had a strong interest in this topic because I live on the Chesapeake Bay and am aware of the effects of the dramatic decline in oysters in recent years on the health of the bay as well as Maryland's economy. I intended to explore patterns in weather, oyster harvesting, water quality, pollution, and economic data in order to better understand the causes and results of the oyster population decline, as well as the success of efforts to restore the oyster population. My hope was to come up with a visualization to provide meaningful insight into this issue that could be leveraged to assist oyster restoration efforts.

Related Work

In the beginning stages of my research I found a number of articles and anecdotes about the importance of oysters in the Bay and in the region that inspired me. Oysters are especially important for filtering harmful nutrients and toxins out of the water in order to provide sufficient conditions for other species to survive in the Bay. One article included a time lapse video of how effectively oysters filter and clean the contaminants that plague modern society out of the water (http://goo.gl/AgbLgT). Having observed how poor the water quality of the Bay has become first hand, I was both fascinated and inspired by the potential, and did quite a bit more research on current oyster restoration efforts into which the Maryland government has invested millions of dollars in recent years.

https://www.youtube.com/embed/VTuBbuUro4g?rel=0

Questions

My initial questions for this project were:

-How has the oyster population changed over time in the Chesapeake Bay?

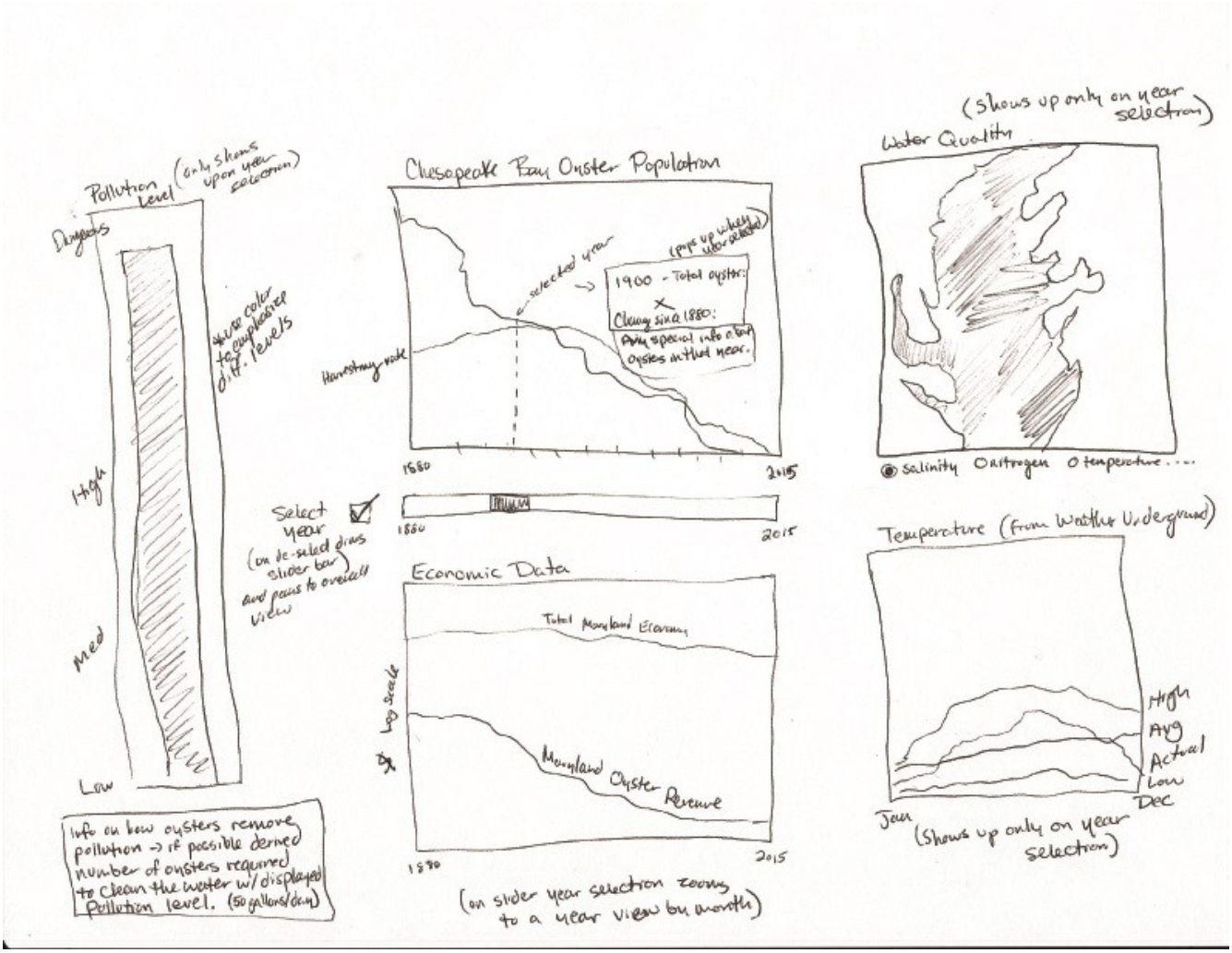
-What are the primary causes of the decline in oyster population?

-What are the effects of the decline in oyster population on the water quality (oysters filter the water), and the local economies?

-How can efforts to reintroduce oysters be most successful? What other elements need to be considered?

Over the course of my work these changed slightly to the following:
-How has the water quality of the Chesapeake Bay changed over time (with a particular focus on nutrients and conditions that are potentially harmful to creatures living in the Bay)?
-Are there observable patterns in the selected measures based on time of year?
-Are there substantial differences in these measures based on geographic location in the Bay? How do the measures at each station compare to overall Bay averages?
-Are any measures correlated with the oyster population?

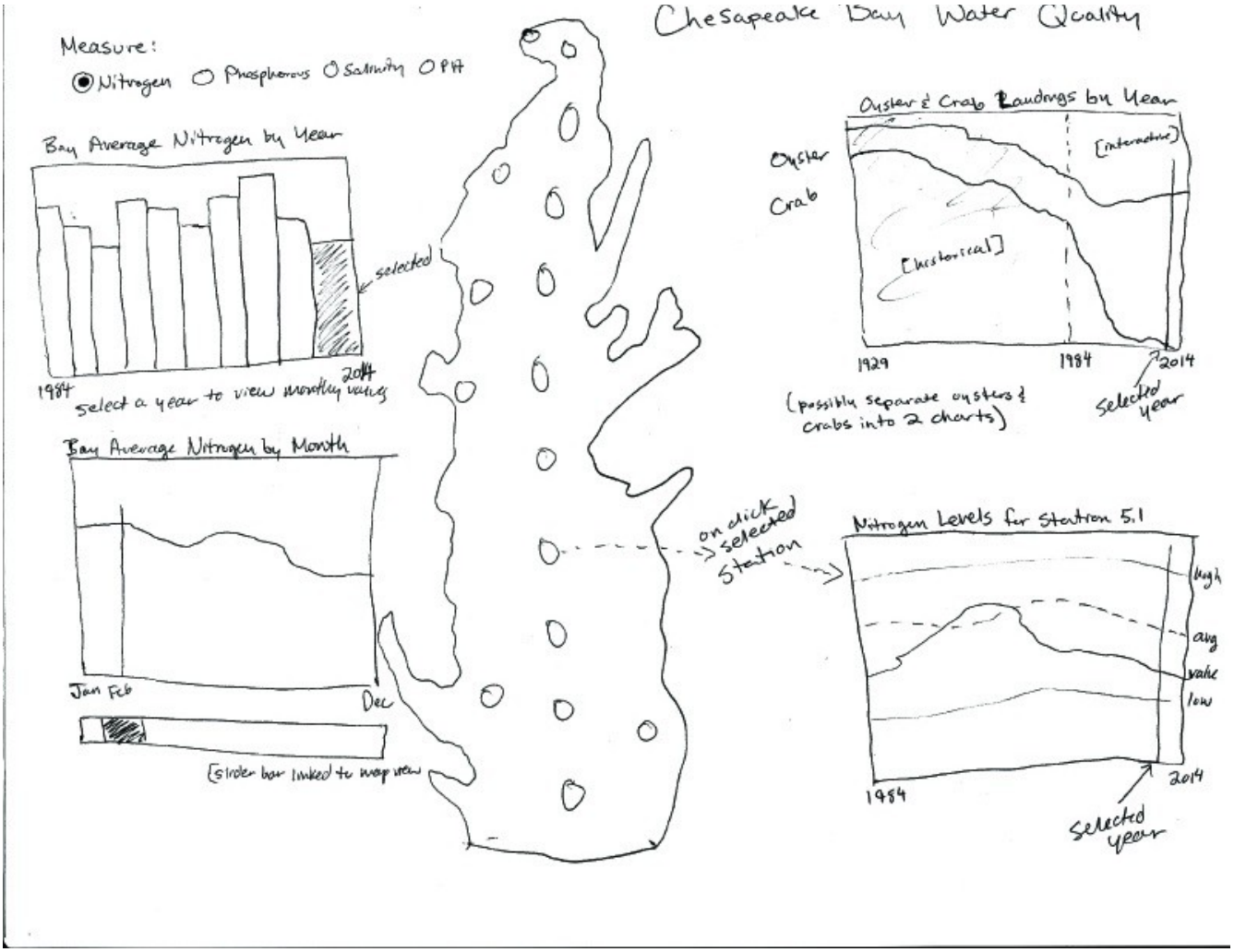
Design
In my initial design I sought to explain through visualization the effect of water quality and pollution on the oyster population and vice versa, as well as the effect on the local economy:



I did decide to modify the topic slightly after some discussion. The most detailed data I had was for water quality from different stations, so instead of having a time series view centered around the oyster population as the focus for the majority of my visualization, I decided instead to move to a geographic layout with the water quality stations as the focus. I thought this would result in a more interesting visualization because the data was much richer.

My modified view centers around a plot of the water quality stations on a map of the Bay coastline. In this design the user can choose a measure (i.e. nitrogen, dissolved oxygen, water temperature, salinity,

water clarity) and a bar chart updates with the average yearly values for all stations across the Bay. The user can then select one of the bars to choose a year to filter down for the data in a chart showing the monthly averages across all stations. The user can then interact via a slider bar representing all months in the selected year, to observe how the values change by month (the current month will be represented by a line on the monthly averages chart, and the size of the circles representing each station on the map will change based on the relative values of the selected measure for that month. Since I did want to keep a focus of the project on the relationship between water quality and oyster populations, if the data supports it I plan to have a chart on the right that shows the oyster population by year, with a line representing the current selected year. Finally, the user will be able to click on any single station and view the full history of the selected measure across all years with available data.



Data

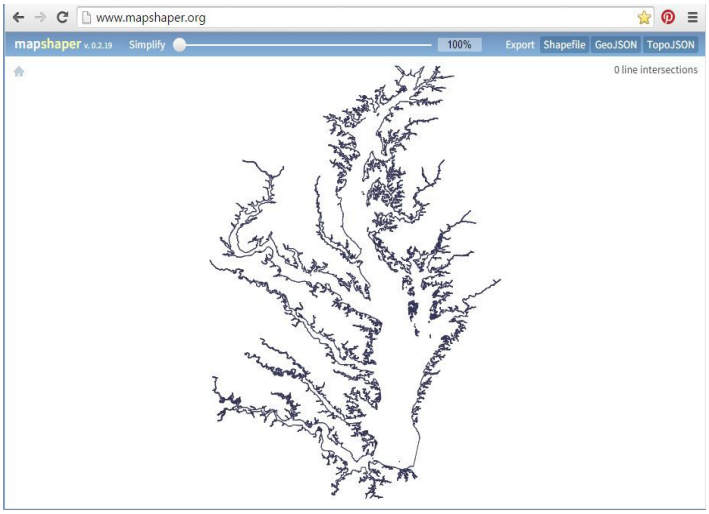
The data gathering phase was far more difficult and time consuming than I expected. The data I needed resided in many different places and forms. I found many charts representing some aspects of the data so I assumed it would be easy to find the source data as well but this was not the case. After thoroughly researching the topic and data for days I still had very little to work with. It got to the point where I thought about switching topics entirely to something with more readily available data. However I feared I would not learn as much from this project about overcoming real world challenges if I took the easy way out and based my topic off of readily available data instead of finding the data I needed to answer the question at hand, so I decided to press on.

After extensive research I was able to find the data I need, collecting it from a broad range of sources. There is an API for Chesapeake Bay water quality data but it is very slow and required pulling data in small batches. The data is also fairly dirty and required a lot of de-duplicating and cleaning. I scraped the oyster and crab population data from HTML tables on web pages. I was also able to obtain water quality data by emailing the Department of Natural Resources and NOAA.

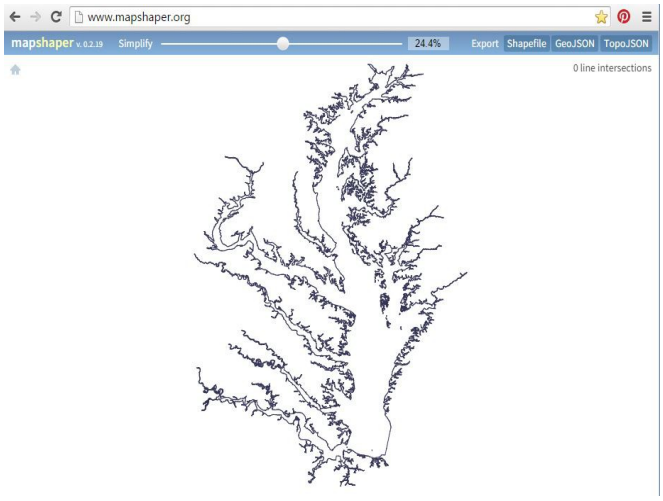
One of the hardest parts of the data to find was the shape file for the geographic visualization. While there was a significant amount of GIS data for the Chesapeake Bay, most of these shape files were created as part of studies and included extraneous information which prevented me from being able to use them. After much research I was able to get a hold of a shape file for the Chesapeake Bay coastline. I had a lot of difficulty trying to convert my shape file to topojson since I have a Windows machine.

After much research on the topic I found the Mapshaper online conversion tool (http://www.mapshaper.org/) to be very helpful. I was able to input a 5.2MB shapefile and convert to a 4.6MB topoJSON file. The tool includes a feature to reduce complexity, and I was able to reduce the file to 726KB with hardly any noticeable visual loss.

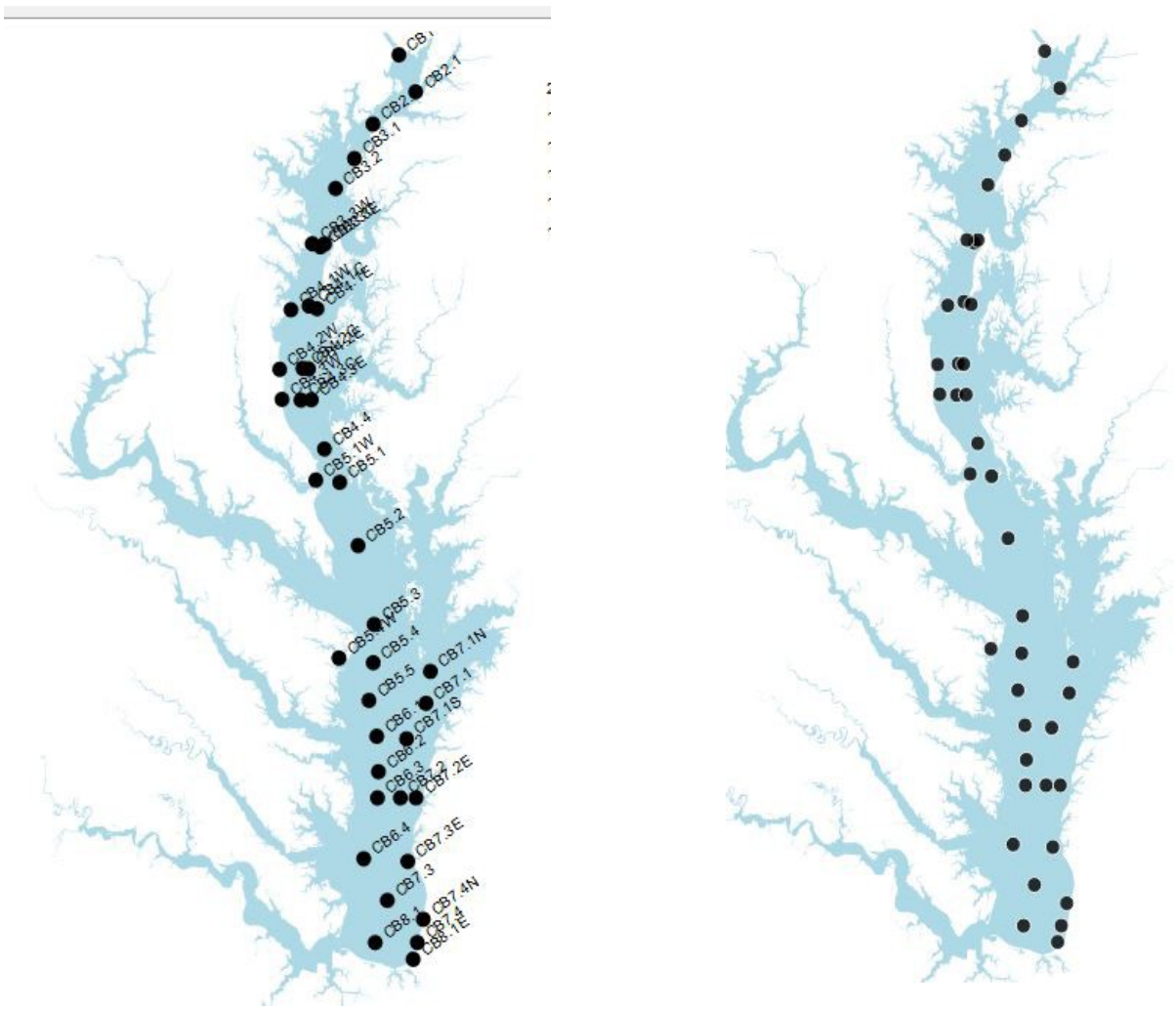
Coastline at 100% (4.6MB):



Coastline at ~24% (726KB):



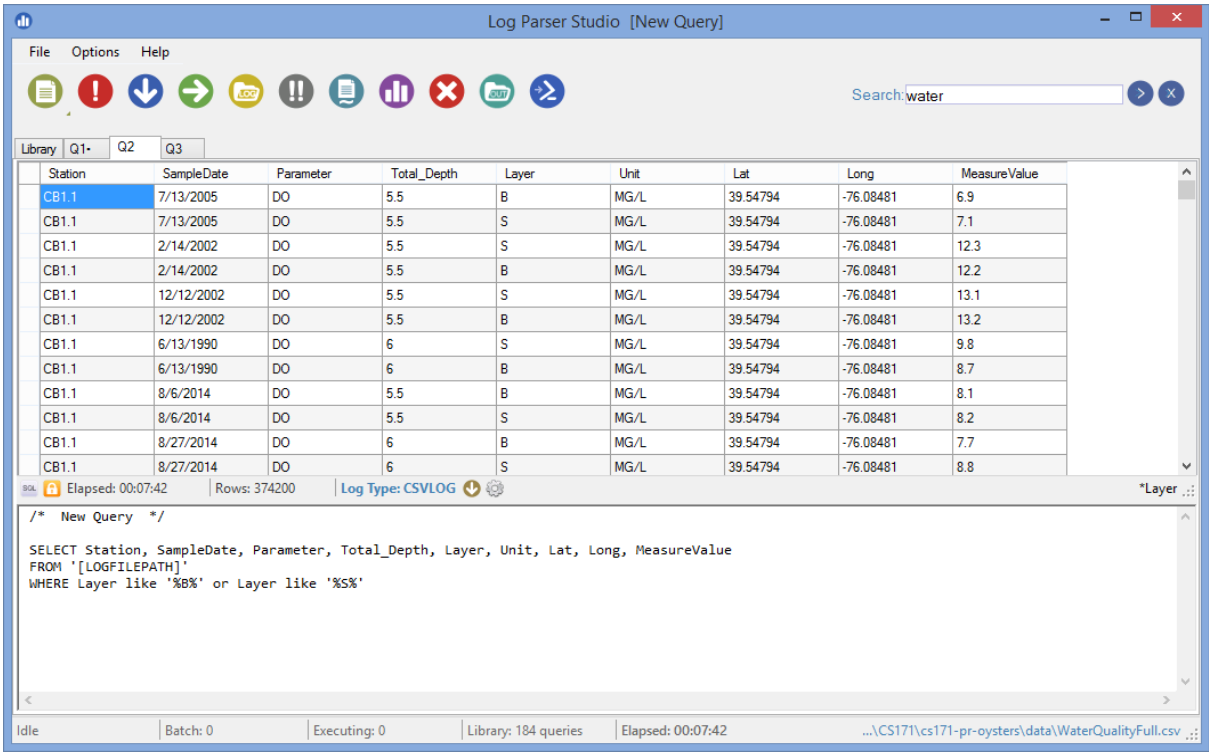
I figured out how to work with my TopoJSON output file via the TopoJSON library. I initially ran into issues applying the projection function to my station coordinates (it was outputting latitude and longitude values of over 50000) but when I put the projections of the coordinates in a transform/translate attribute element the projections were corrected. I started off showing the labels for all stations but then decided that this cluttered the visualization too much, and changed it to only show the station label (and highlight the station by changing the fill and text to red) when the user mouses over one of the stations.



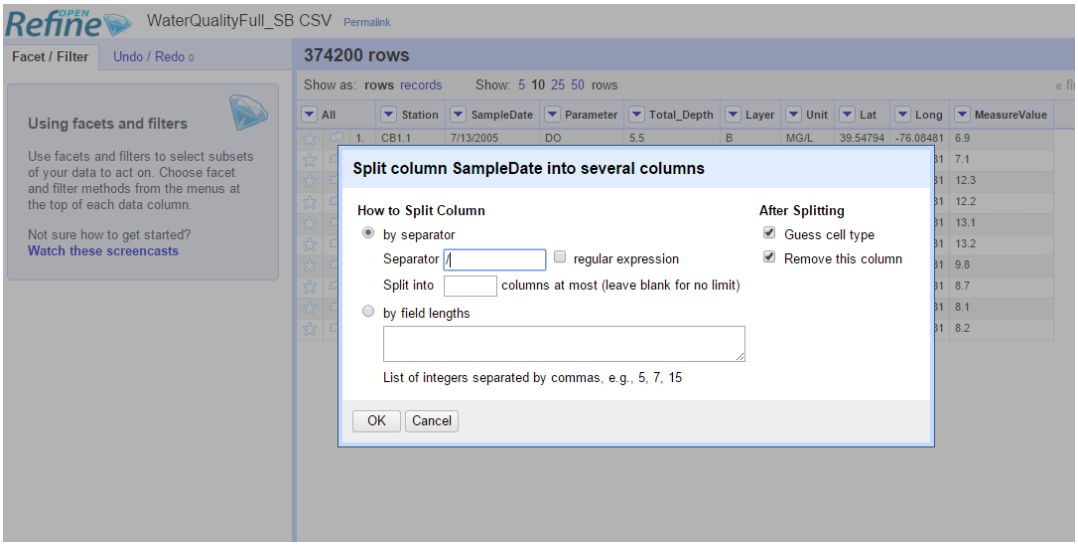
Chesapeake Bay coast shapefile implemented in D3/topoJSON with plotted stations shown with station labels displayed (left), and station labels only displayed on mouseover (right).

Exploratory Data Analysis

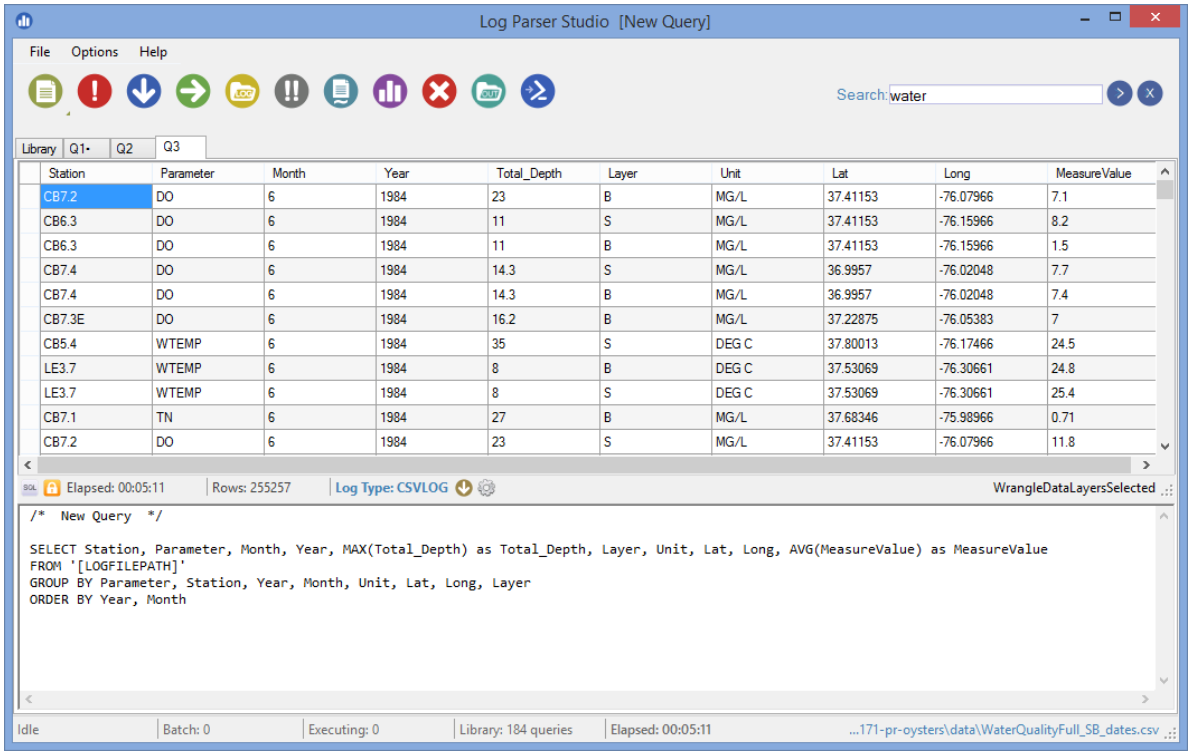
I built most of the visualization on a small sample of the data. When expanding to a larger sample of the data set I ran into issues with some stations reporting multiple different measurements for the same value for the same date. I decided that the best course of action would be to average these values since the samples were taken on the same day with all other parameters being the same. I tried using OpenRefine first but found that it couldn't handle the size of my dataset (about 1.5 million rows). I then found Microsoft's Log Parser tool which enabled me to remove all measurement depths I was not using (I decided to only use surface and bottom measurements in the visualization), and quickly and easily clean the data and average the multiples.



After removing duplicates and the values I did not need I was able to load the file (which was down to 374,200 rows) into OpenRefine to parse the date text string into the year and month columns I needed.

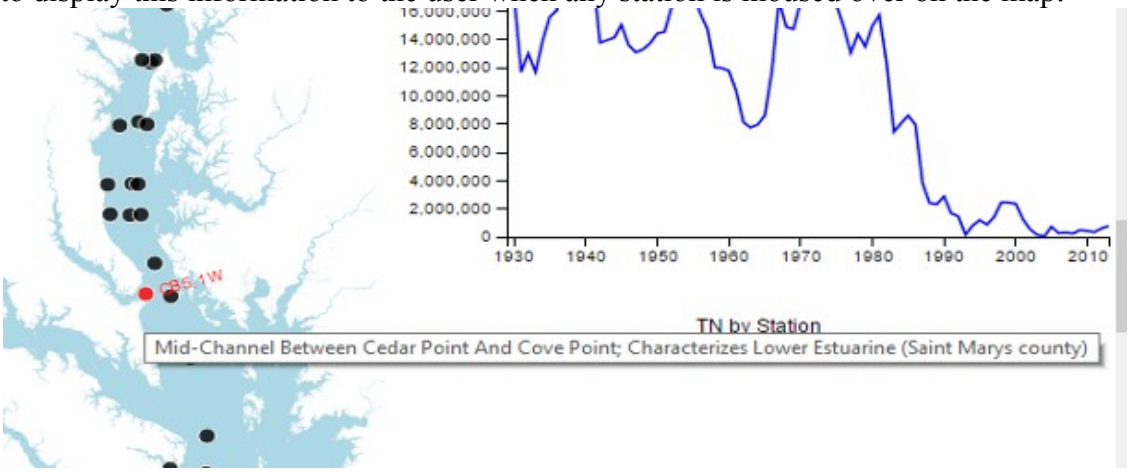


Finally, I loaded the file back into Log Parser and executed a SQL query against it to select and group the columns of interest.

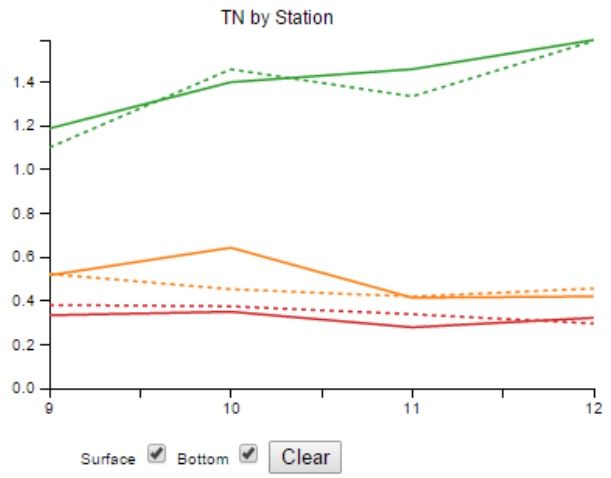


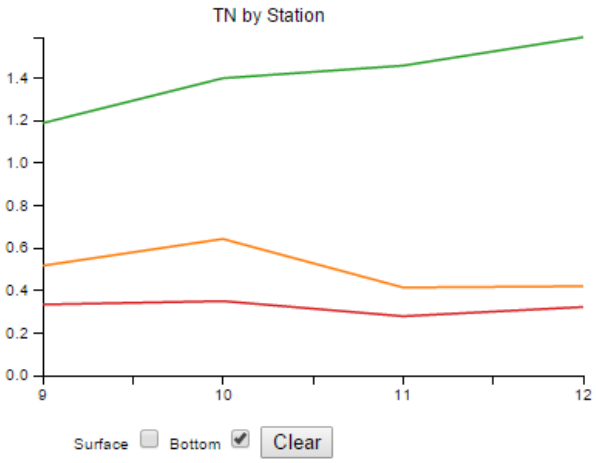
Design Evolution and Implementation

Since the Station Information data had fields for the description and county I decided it would be useful to include this information so that if a user found a particular station to be of interest based on use of the visualization they would have enough information to research that area/county further. I added a tool tip to display this information to the user when any station is moused over on the map:

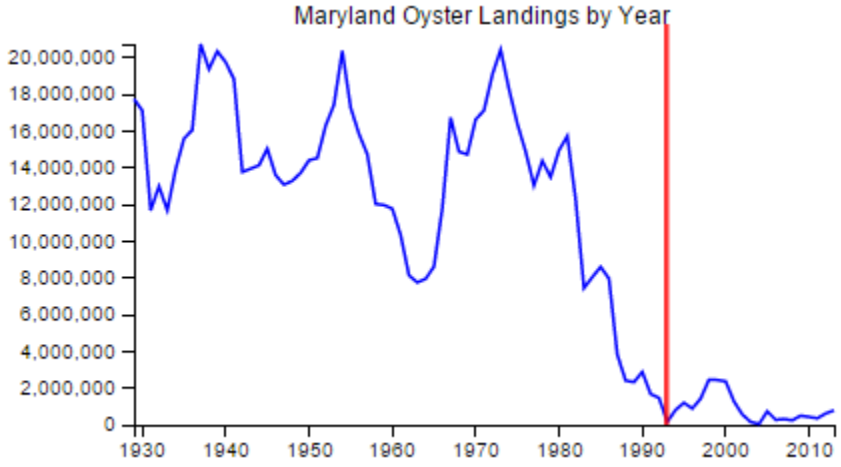


I had originally intended to display a detailed view of the selected measure over the available data time range for a single station at a time (both for the surface and bottom measures) along with the high, low, and average values for all stations. However, during implementation I realized that it would be much more valuable for the user to be able to dynamically compare that values for two or more stations at a time than it would be to see overall averages. I changed the graph to enable the user to add paths for as many stations as desired. Since this view tends to get crowded quickly I also added a clear button, and checkboxes to show or hide either the surface or bottom measures.



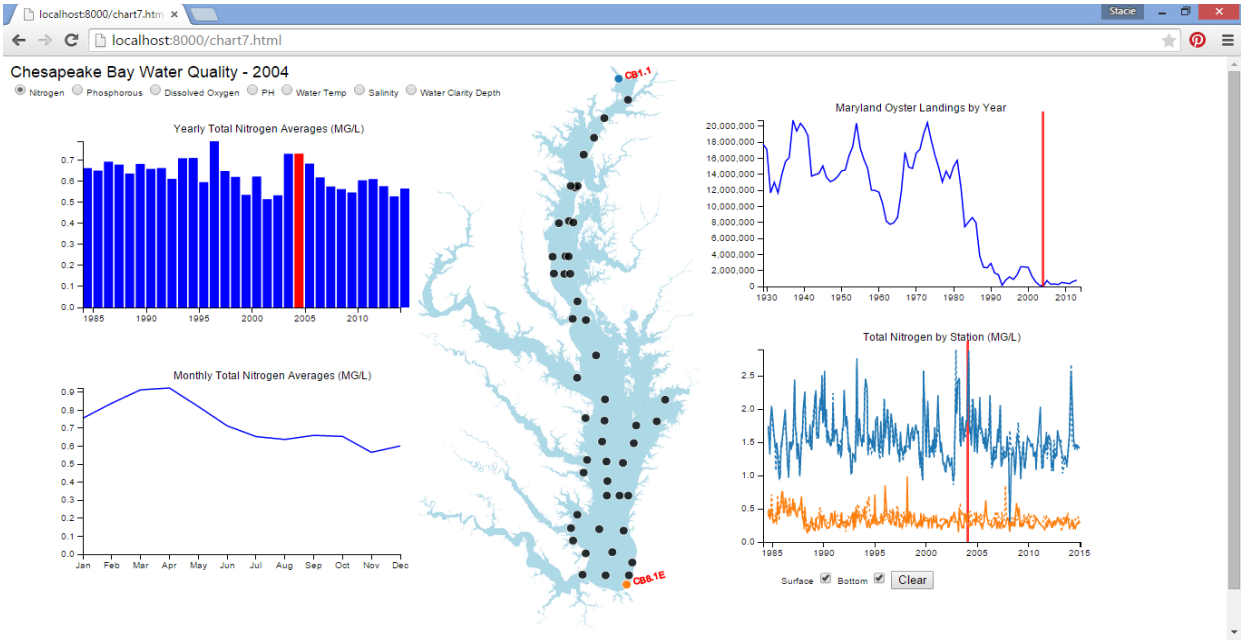


In my original design I planned to use slider bars to enable the user to select a year and month. On the charts with yearly data I drew a red line to show which year was currently selected. I then decided it would look cleaner to enable the user to drag this line directly on the chart for the oyster population line chart, just as the design already enabled them to select a year on the Yearly Averages chart by clicking on of the bars. I made the decision to implement year selection in this way because I thought it would better support analysis to let the user select a year to apply to the entire visualization directly from the chart they are viewing. This would allow for something like dragging the bar to a spike or decline in the oyster population and then having the other charts update to show the corresponding water quality values.

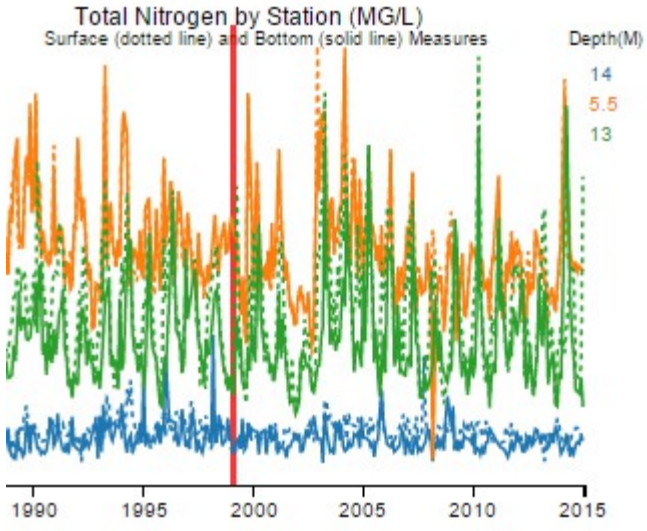


Originally I had all of the graphs set up to change as I dragged the slider bar on the oyster chart but this caused too much of a lag due to the size of the data so I changed it to only modify the other charts on the drag end function. I also decided to change the measures by station view so that it shows all data instead of just selected year. I thought that having all info would enable better trend identification and comparison to the oyster landing data. Since this introduced a lot of data into the small graph I made the decision to enable zooming on this graph as well so that the user can look at monthly detail in addition to the big picture.

I decided against sizing the stations on the map by values as I had originally planned because it looked way too cluttered, and was redundant anyway so it did not add much value. Instead, I color coded the selected stations to match the colors on the station comparison graph, which very effectively enables the user to visually connect the series on the station comparison graph to the geographic representation of the station.

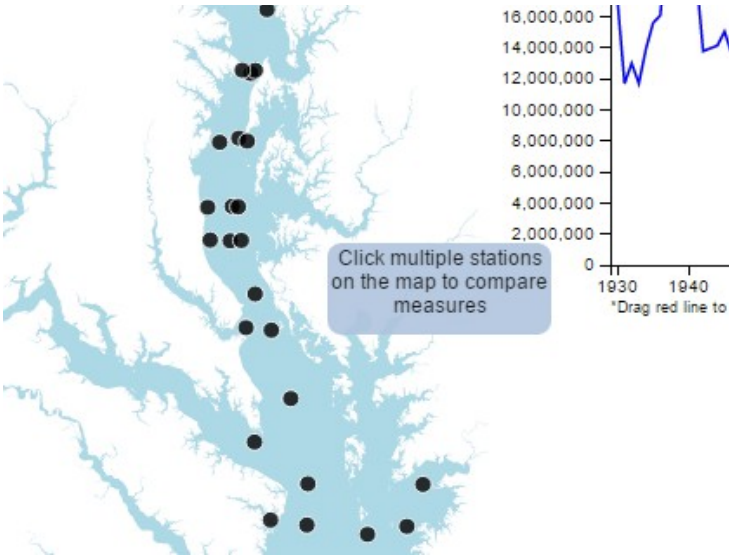


During my implementation of the station measures graph I found myself wondering whether water depth had an impact on measures. In order to enable users to factor this in I added color coded values to the right of the station measure comparison chart to show the depth at each of the selected stations.

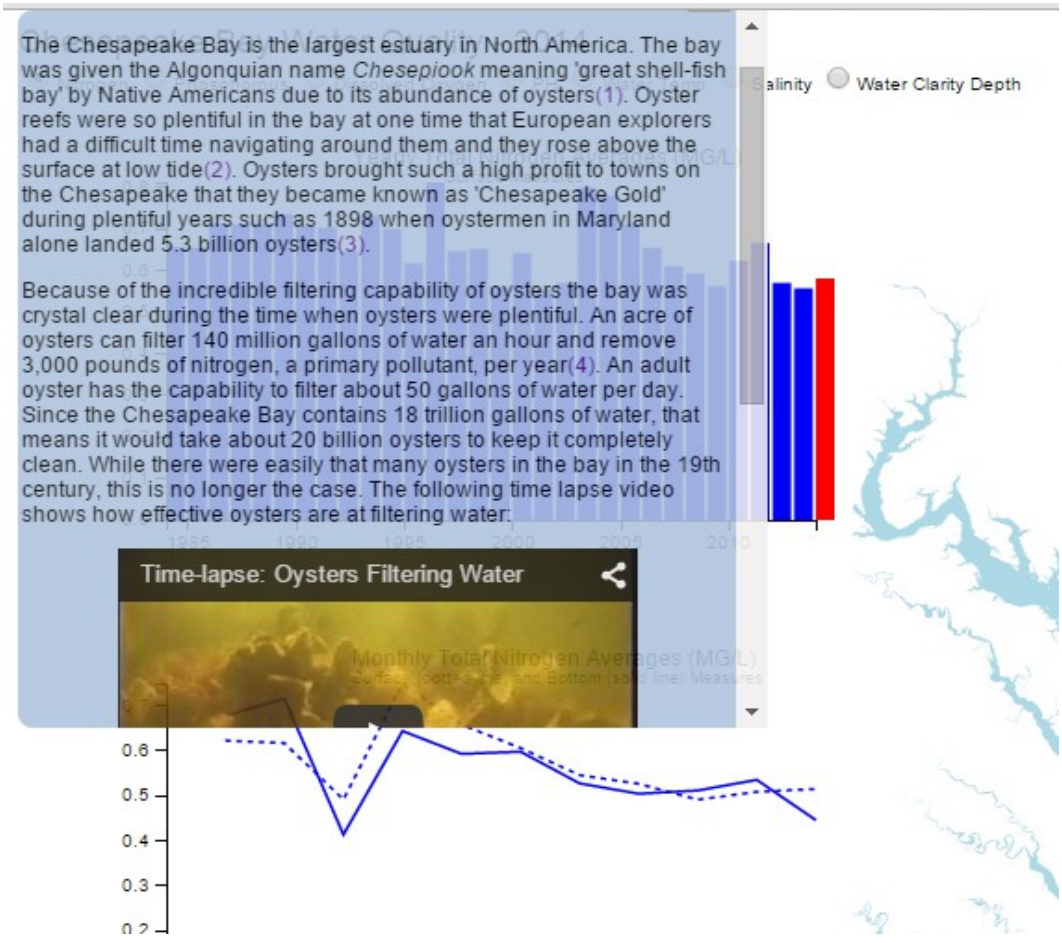


Since I added a lot of interactive features I decided to add a tutorial to introduce the users to all features when they click the tutorial button. I chose to make the tutorial interactive by having boxes with instructions pop up sequentially in the areas of the features I wanted to show, and go away once the

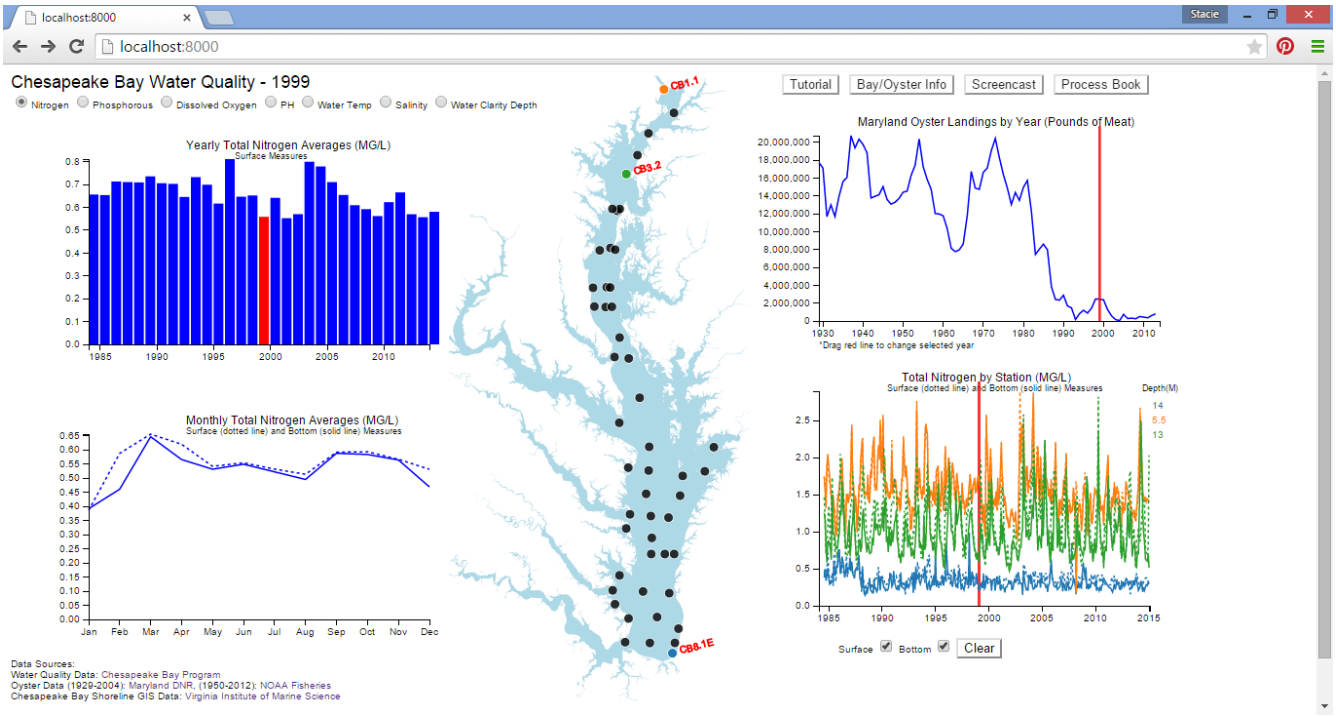
user engages that feature.



Finally, since I wanted to provide some history and background to convey the importance behind the topic of my project, I decided to include an information slide detailing some interesting facts related to Chesapeake Bay history and the role of oysters in keeping the Bay healthy. I embedded the time lapse video that was one of my original inspirations for this project, and provided a link to the Save the Bay foundation for users who are driven to learn more.

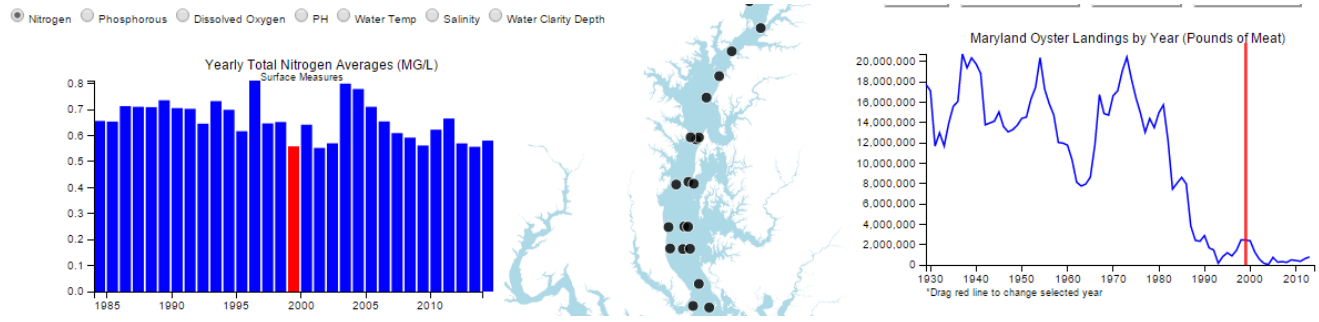


The final implementation of my visualization:

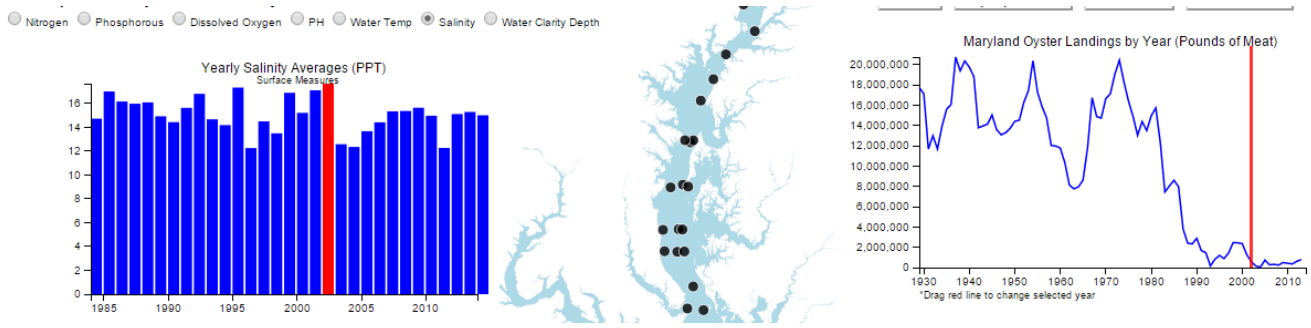


Evaluation

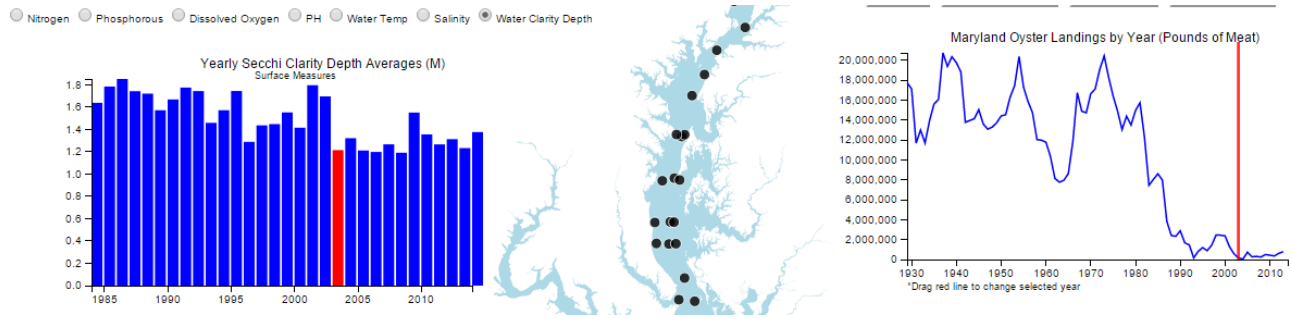
Oysters are known to filter pollutants out of the water, but in my research I did not find any actual experiments comparing water quality data to the oyster population to determine how effective they are. Using my visualization I was able to prove this fact as there is a clear correlation between the years when the oyster population increased and decreased, and the levels of pollutants as well as the water clarity. In 1999 the oyster population hit a relative peak before sharply dropping to a low in 2003. This is reflected in the nitrogen yearly averages chart, as there was a big dip in the nitrogen levels in 1999 and a sharp increase again in 2003. This proves that oysters really do have a significant effect on nitrogen levels. In fact the nitrogen yearly averages chart loosely resembles the inverse shape of the oyster landings chart.



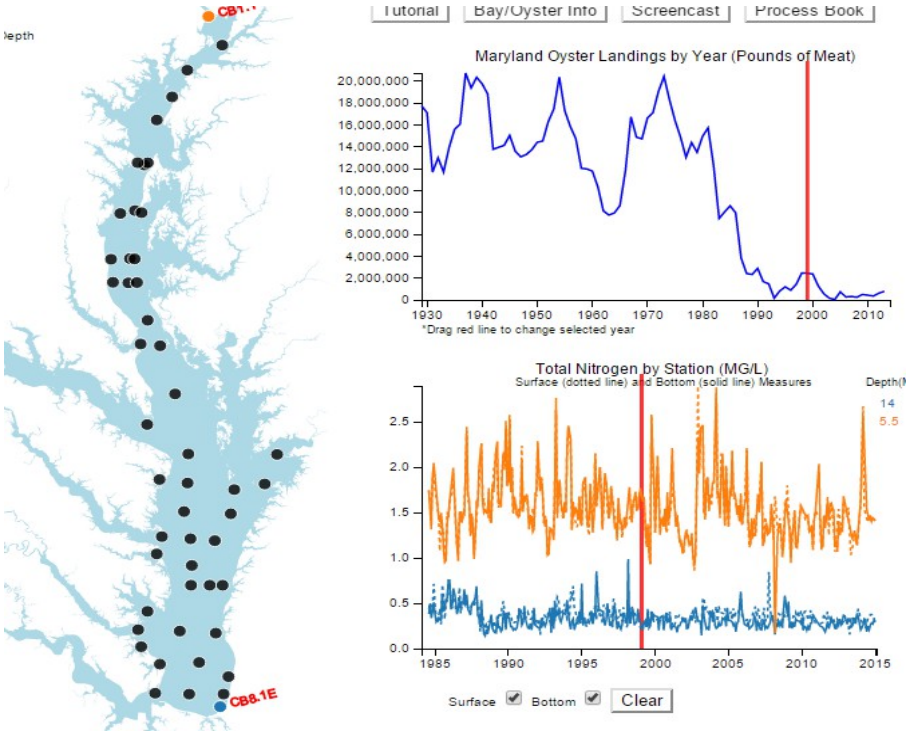
An interesting discovery I came across was that salinity also seemed to follow the same pattern as the oyster population but with a bit of a lag. It would be interesting to do more research to see if the water salinity contributes to health of oysters and/or their filtering capacity.



As expected, the water clarity has gone down with the drop in oysters. The effect of oysters' filtering capabilities on water clarity has been well established, but it was interesting to see this reflected in the data. The water visibility went down from 1.8 meters in the 1980s to about 1.2 meters in the early 2000s, and then slightly back up as the oyster restoration efforts have been taking hold and oysters slightly increasing.



Another interesting fact I discovered from my visualization is that the nitrogen levels are significantly higher at the north end of the Chesapeake Bay than they are near the southern end where it meets the ocean. It would be interesting to research more to determine why this is, whether it is due to salinity differences, depth, or proximity to major cities.



Overall, I think my visualization works very well, and is effective in conveying the intended information and telling the story of how the oyster population relates to water quality in the Chesapeake Bay. To further improve it I would like to add water quality data for each of the tributaries and see how those measures compare to the measures in the main part of the Bay. I would also like to add in data for crab landings and see if the crab population has any correlation with the selected measures.