INITIAL PROPOSAL:

Create a proposal document addressing the following points. Use the points as headers in your document. Each team member must push this document into his/her homework directory. As a ballpark number: your proposal should contain about 3-4 pages of text, plus 5-6 pages of sketches.

Basic Info. The project title, your names, e-mail addresses, GitHub ids, a link to the project repository (I encourage you to make it a public repo).

Title: something along the lines of "Influenza …."

Jacob Moon:

Email: jamoon@wpi.edu

Github id: Yeknomo

Virginia Nunez Mir:

Email: vcnunez@wpi.edu

Github id: vcnunezmir

GitHub repo: https://github.com/vcnunezmir/DataVisInfluenza

Background and Motivation. Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

We are both bioinformatic students, so we knew we wanted to work on something bio related. Influenza is such a widespread disease that there is a wealth of information out there to be explored. Furthermore, the database on the influenza virus shows connections between many species, and we found a lot of interest, as well as potential, in it.

Project Objectives. Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits.

What are the connections between the locations of certain birds known to transmit influenza, and known human influenza patients?

Are there any noticeable connections between strains and their characteristics?

Are there any noticeable connections between the spread of influenza and the characteristics of the spreading strains?

Data. From where and how are you collecting your data? If appropriate, provide a link to your data sources.

Datasets: https://www.fludb.org/brc/search_landing.spg?decorator=influenza

Most, if not all of our data is going to come from the Influenza Research Database. The page includes many different datasets, including animal surveillance data, data on influenza strains, and data on human isolates of the influenza virus. Our visualizations will be composed of combinations of the different datasets found in this site. For example, one of the graphs might use just the information from the human isolates dataset, comparing one attribute to another. A different graph might compare one attribute from the human isolates to the same attribute in the animal surveillance dataset. Yet another graph could compare attributes from the human isolate dataset, such as strain, to information obtained about the strain from the strain dataset.

Mapping the states: http://eric.clst.org/Stuff/USGeoJSON

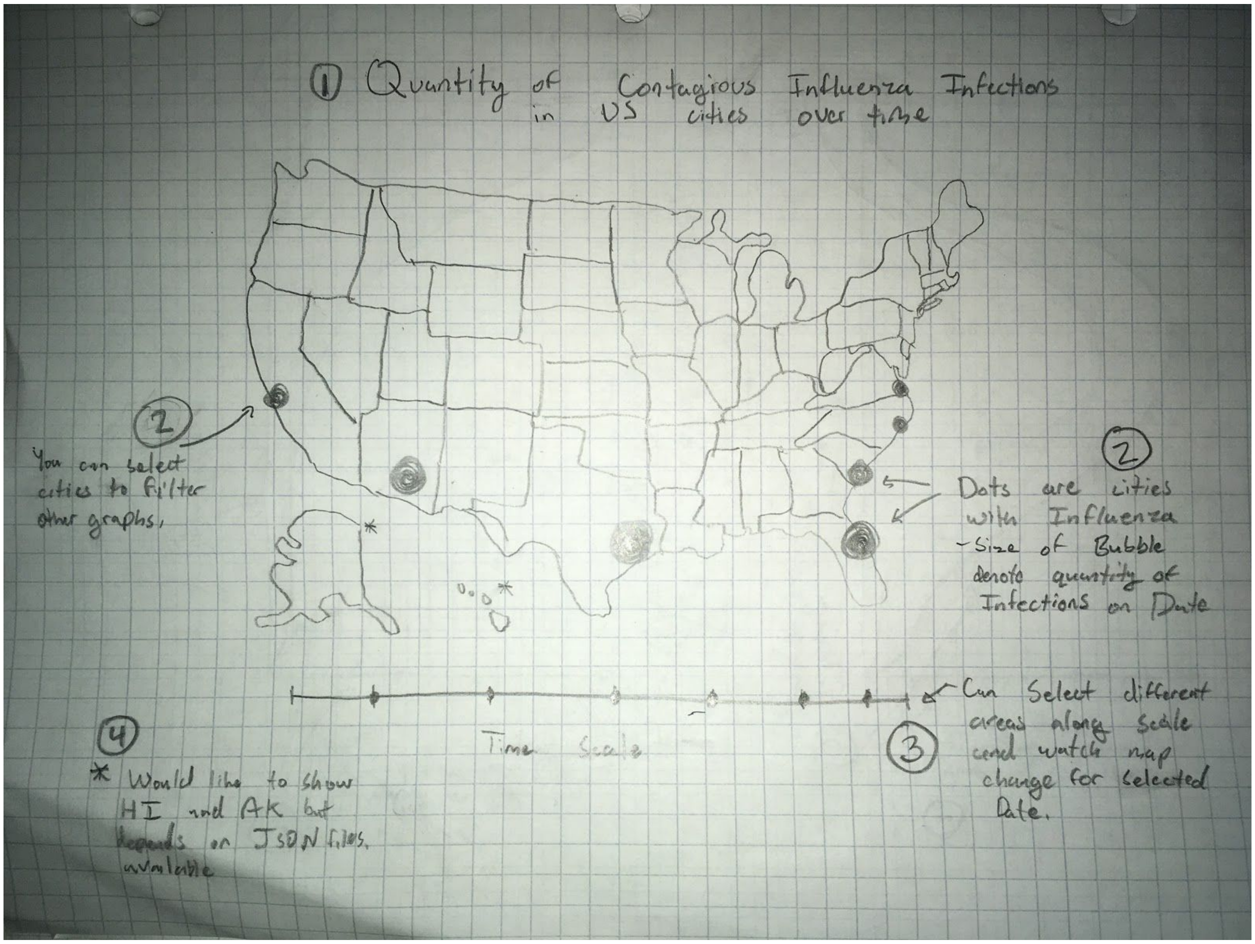
This is a GeoJSON file for constructing a map of the Us with all of the States in it. It will also be used to create a projection of all latitudes and longitudes on the SVG, this will then be used to incorporate cities into our mapping.

Data Processing. Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?

Due to all of our data being spread across different datasets, a necessary aspect of our project will be to sift through all of the different datasets and compile the important information from each. There probably will not be any calculations involved during data processing, considering most of the data we are interesting is qualitative and categorical. However, any calculations will probably be made in excel before importing the file into d3.

Visualization Design. How will you display your data? Provide some general ideas that you have for the visualization design. Develop three alternative prototype designs for your visualization. Create one final design that incorporates the best of your three designs. Describe your designs and justify your choices of visual encodings. We recommend you use the Five Design Sheet Methodology.

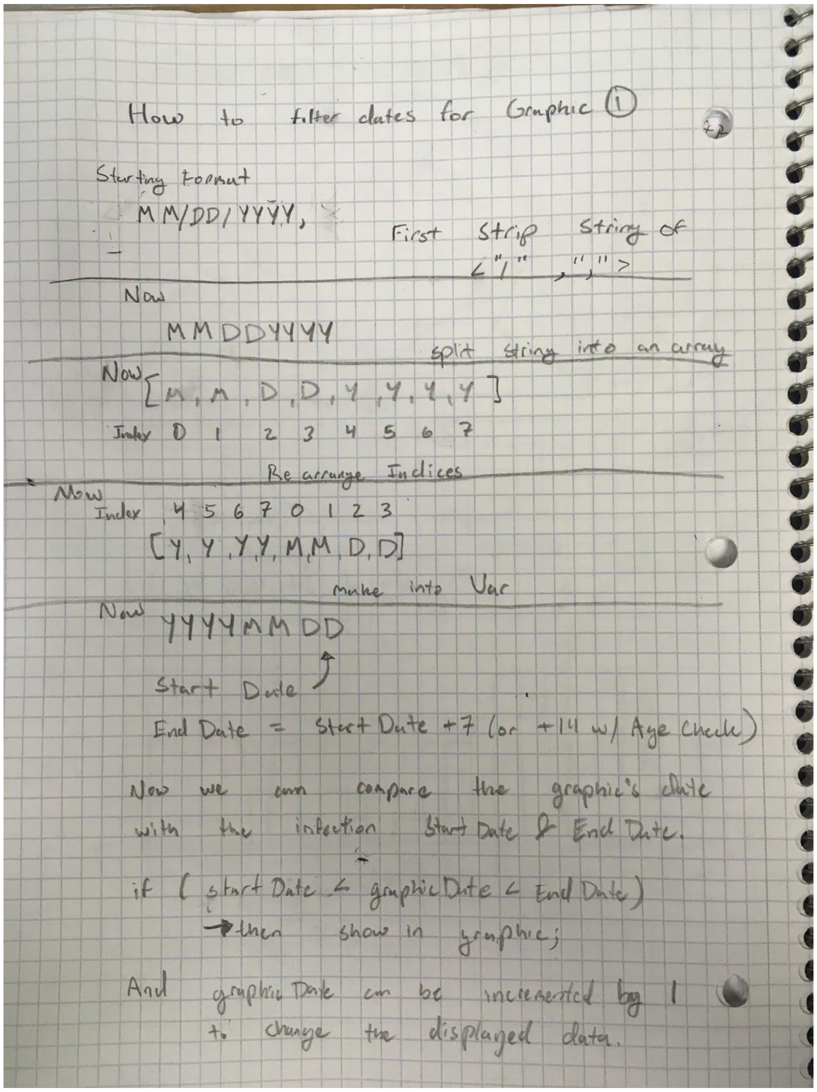
This is the organization of the visualization page of our program. It will feature 5 major visualizations which will each be covered below. Each graphic will have it's own SVBG container and may be linked with the other graphics, to work together.


① Quantity of Contagious Influenza Infections in US cities over time
②
You can select cities to filter other graphs.
②
Dots are cities with Influenza
- Size of Bubble denote quantity of Infections on Date
Can Select different areas along scale and watch map change for selected Date.
③
Time Scale
④
* Would like to show HI and AK but depends on JSON files available
*
*

This is the main graphic for the project it is a map of the US with the different states. (1) The map will have each the states as a default tan color or a red color for infected states, as time goes by circles are added to the map to represent active contagious infections. An active contagious infection is either seven days after diagnosis or 14 days if the person is a child or elderly. (2) The size of the circles will display the quantity of infections and will be placed on the cities the infections take place in. The circles can be selected to filter the other four graphs by city that was selected. (3) On the bottom on the chart will will be a time selector, will have increments of months and a play button for the program to iterate through the days of the month. The goal of this map will be to show the flu "travel" throughout the US as time goes by populating across cities in a wave.

How to make this graphic:

We will first need to make a dataset of all infections, containing Latitude, Longitude, Start of Infection date, End of infection date, City name, state, and Strain Subtype. Then we will set a start time for the graphic, which will have no infections. Upon starting the program, it will iterate through a new day each second. Or the user can select a month on the graphic and it will display the flu infections for the first day of the month. Then the graphic can be played from that day on. Calculating and iterating dates may be a challenge so to simplify it, we are going to convert the dates to numbers then increment and compare to filter the dataset. This process is explained in the following notes.



How to filter dates for Graphic ①

Starting Format
MM/DD/YYYY,

First Strip String of
<"/" , "," >

Now
MMDDYYYY

split string into an array

Now [M, M, D, D, Y, Y, Y, Y]
Index 0 1 2 3 4 5 6 7

Re arrange Indices

Now
Index 4 5 6 7 0 1 2 3
[Y, Y, Y, Y, M, M, D, D]

make into Var

Now YYYYMMDD

Start Date

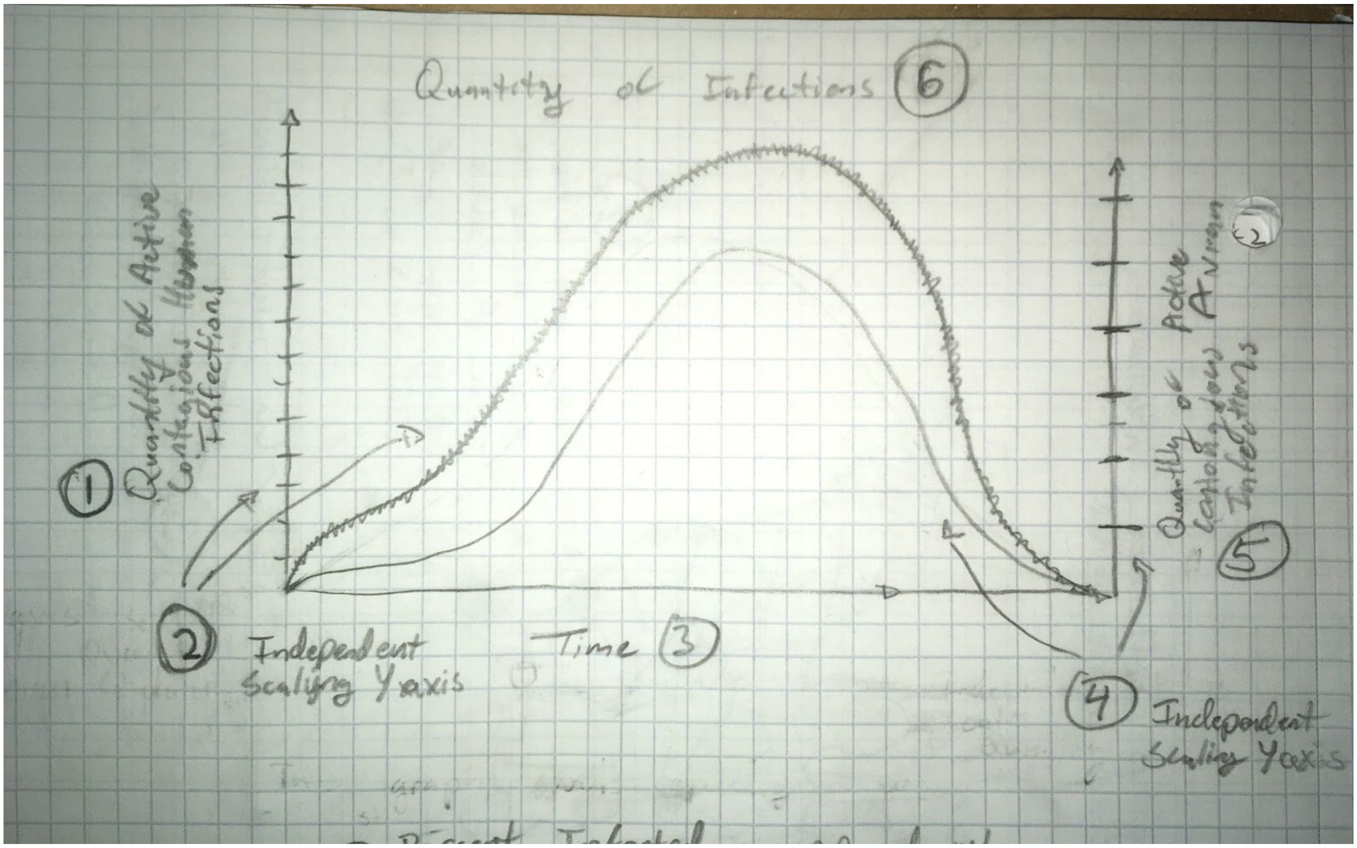
End Date = Start Date +7 (or +14 w/ Age Check)

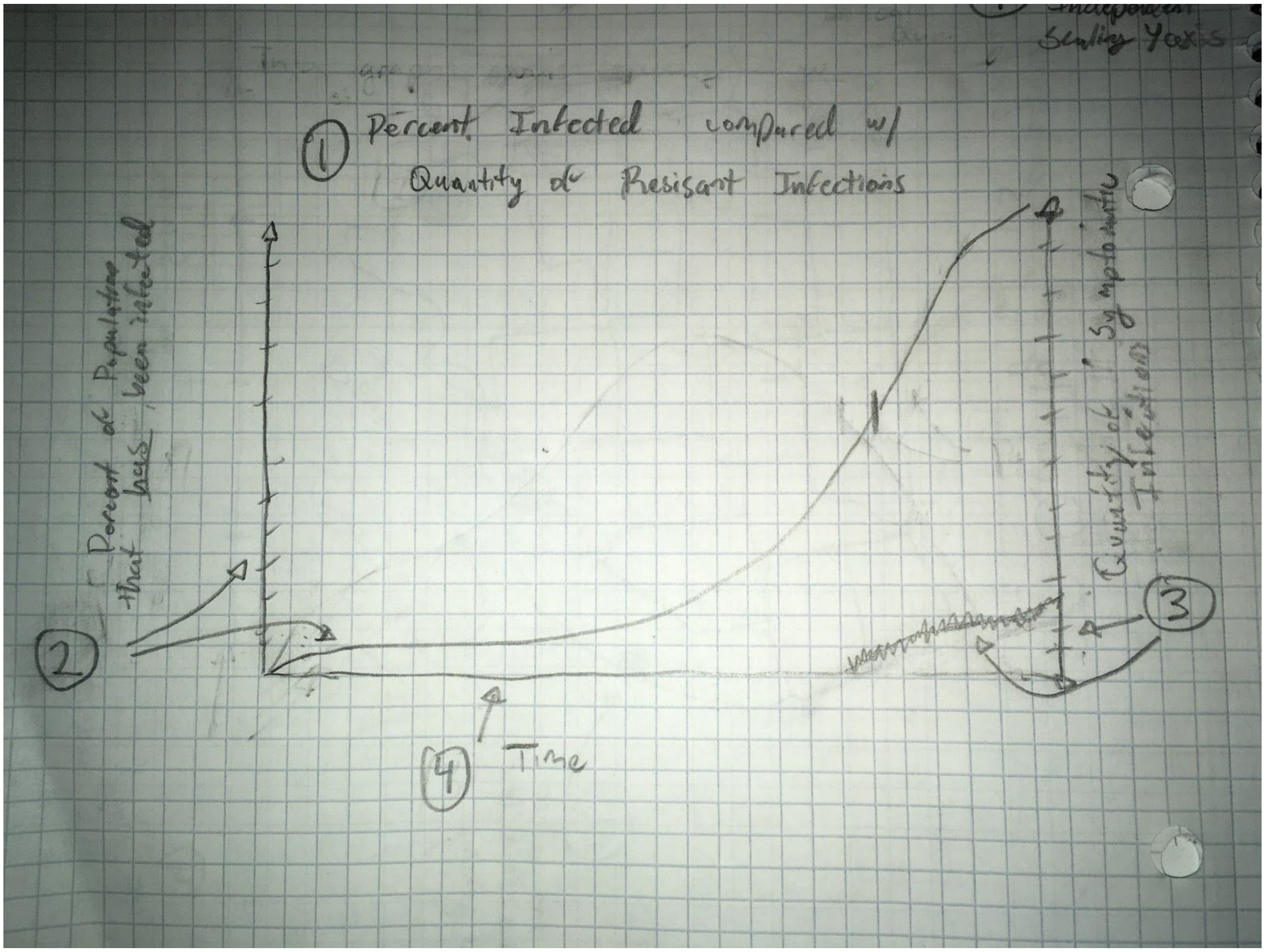
Now we can compare the graphic's date with the infection Start Date & End Date.

```
if ( start Date < graphicDate < End Date)
    → then show in graphic;
```

And graphic Date can be incremented by 1 to change the displayed data.

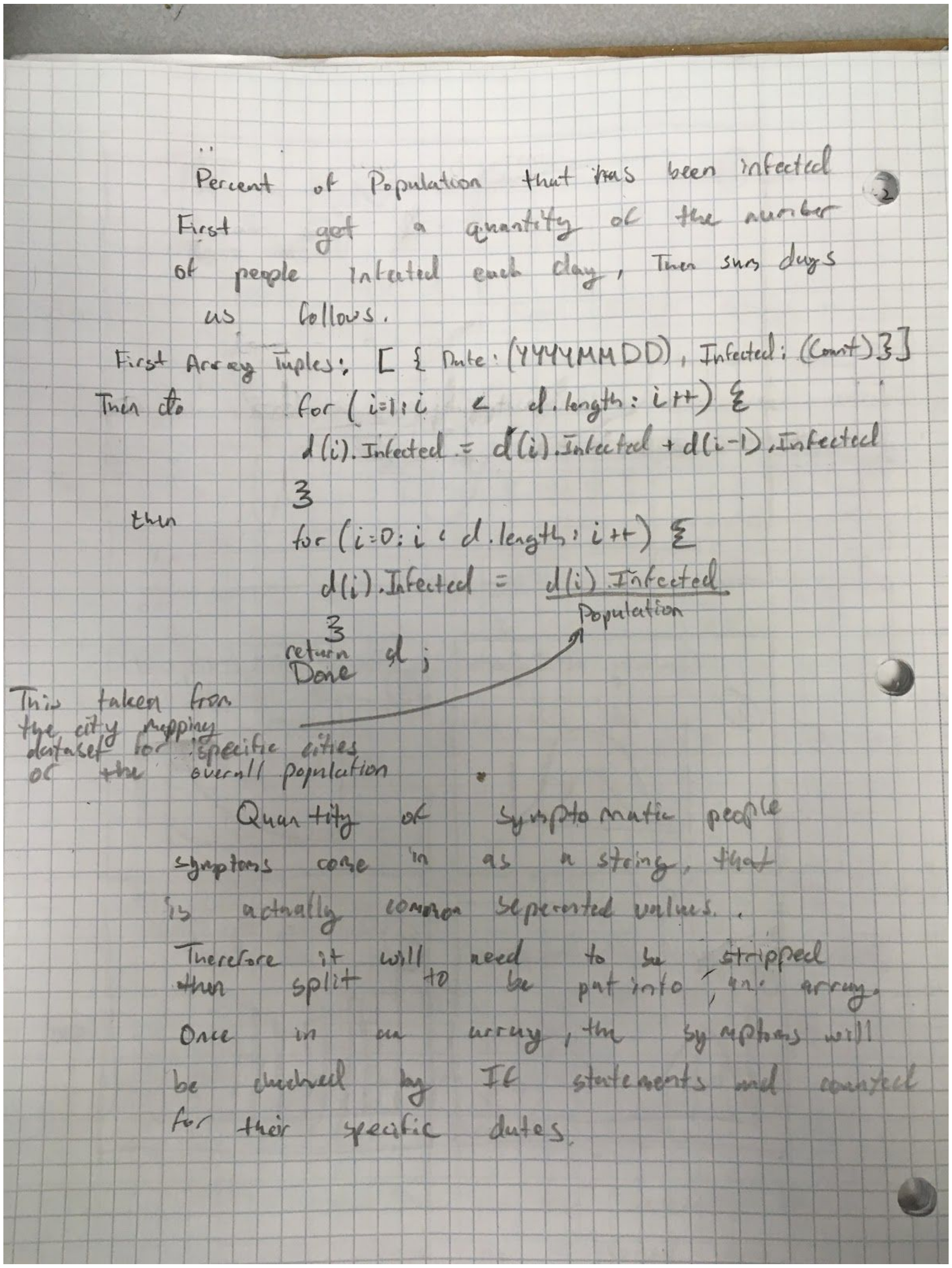
This is the second graphic for our program. It supposed to show the quantity of human infections over time on one line and quantity of animal infections overtime on the second line. This will be used to ho show a correlation between the animal infection rate and the human infection rate. To produce this graphic all we will need to is create a dataset for humans and a second dataset for animals each containing dates and quantities of active infections for the dates.





cThis is the third graphic for our program. It is supposed to show the percent of the city population that has been infected as one line, and another line displaying quantity of symptomatic infections. Both of these lines will be displayed over time. This graphic is intended to show that as a higher percentage of the population is infected, those that are infected are immunocompromised and therefore have a higher symptom rate. To produce this graph, we will need two datasets, one with percent of population that has been infected and dates, and the other with dates and counts of those with symptoms.

This graph is going to require a lot of filtering and therefore is explained in the following notes.



Percent of Population that has been infected

First get a quantity of the number of people Infected each day, Then sum days as follows.

First Array Tuples: [{ Date: (YYYYMMDD), Infected: (Count) }]

```
Then do    for ( i=1; i < d.length: i++) {
               d(i).Infected = d(i).Infected + d(i-1).Infected
           }
then       for (i=0; i < d.length; i++) {
               d(i).Infected = $\frac{d(i).Infected}{Population}$
           }
           return d;
           Done
```

This taken from the city mapping dataset for specific cities or ~~the~~ overall population

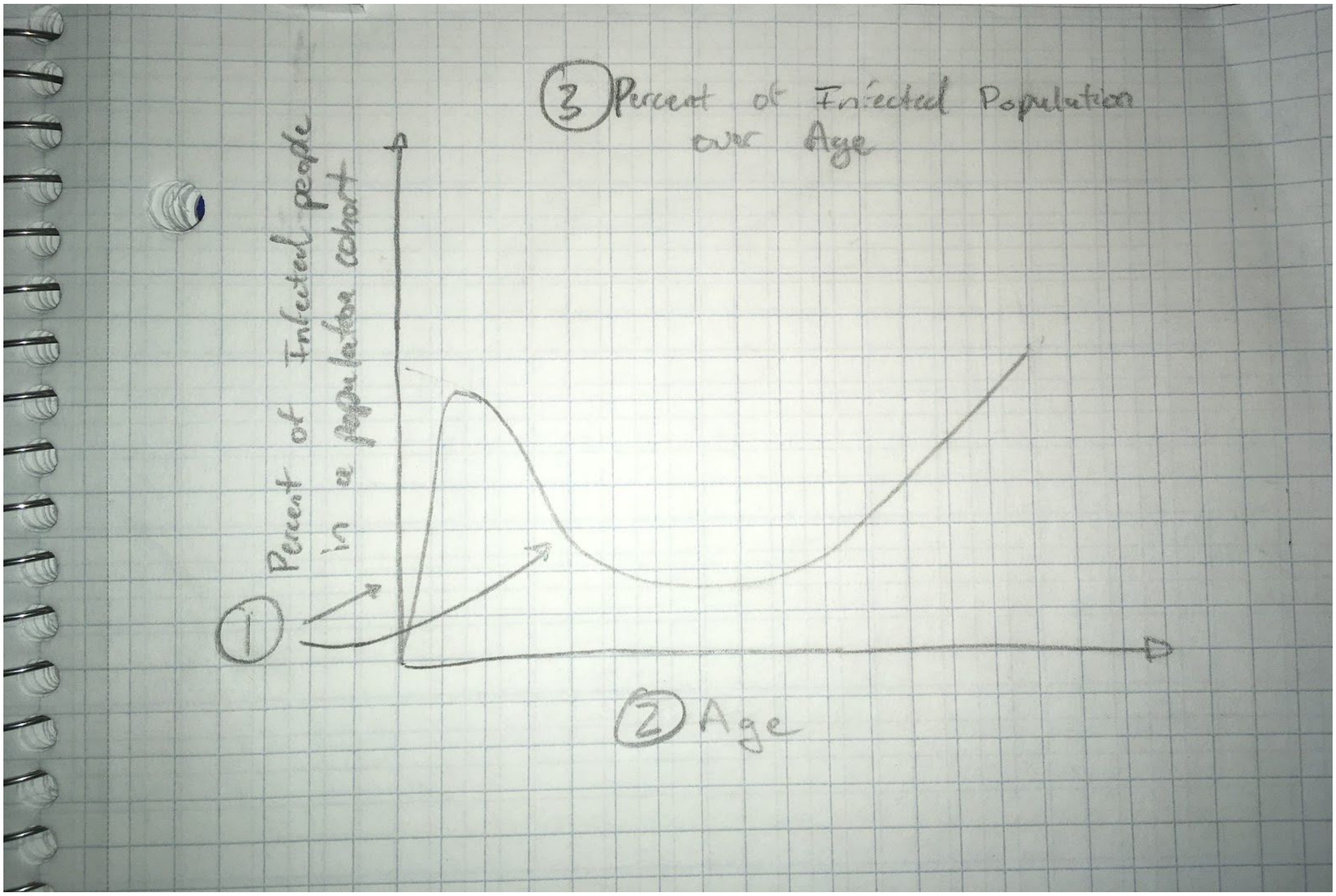
Quantity of symptomatic people

symptoms come in as a string, that is actually common seperated values.

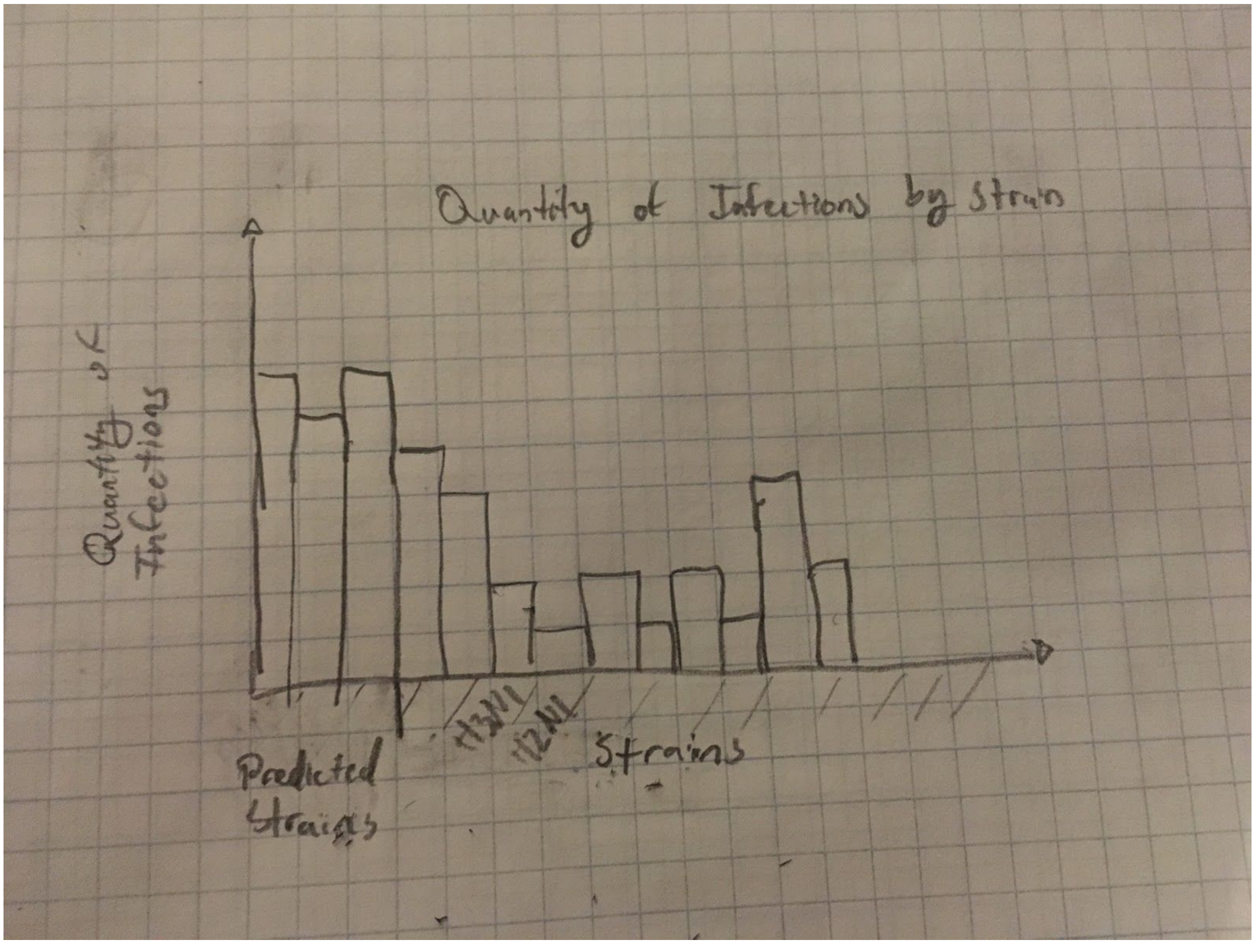
Therefore it will need to be stripped then split to be put into an array.

Once in an array, the symptoms will be checked by If statements and counted for their specific dates.

This is the fourth graphic in our program. It is meant to show a line graph of infected population percent of an age cohort over age. Meaning, we will use an average us census percentage for what percent of the population each age cohort is. Then using we will take those percents and multiply them by the population, to receive the population of each age cohort in a selected grouping (city, state, overall). Then we count infected population of an age and divide the counts by our estimated age cohort population. Therefore, the graph is what percentage of each age group got infected. This will show that higher percentages of immunocompromised populations (elderly and children) become infected with the flu.



This is the fifth graphic for our program. It will show the quantity of those infected by strain. We will highlight and order all of the predicted strains first. We will use the CDC annual strain prediction for producing the vaccines for these. The goal of this is to show what strains actually infect the most people, and how well the CDC prediction holds up, or is affected by the vaccine deployment. This graph will be relatively easy to produce because all we need to do is count the number of infections by strain.



Must-Have Features. List the features without which you would consider your project to be a failure.

Interactivity in our graphs that gives you more information about the specific cases. A demonstration of the cases along a timescale. A map of the location of the cases.

Optional Features. List the features which you consider to be nice to have, but not critical.

Mapping of the birds' locations. Having the graphs update based on what you are doing with each of them (similar to Assignment 4).

Project Schedule. Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

Week of Feb 13: Have general outline of code written out

Week of Feb 20: Have a working prototype

Week of Feb 27: Make changes and final tweaks and improvements

PROCESS BOOK

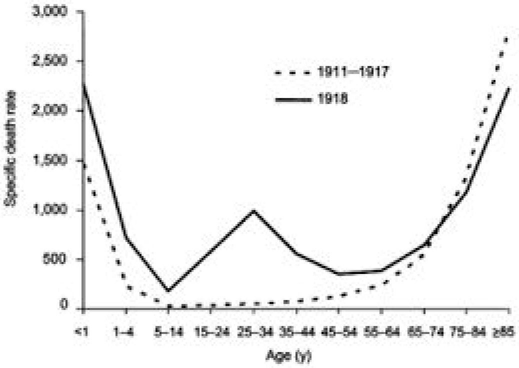
Overview and Motivation: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

For our project we wanted to look at Influenza data and look at when, how and who it spreads to. In the United States, influenza kills more than 36,000 people and more 20,000 are hospitalized each year. However, this is with a vaccine produced each year to fight the predicted strains for the season. While vaccines exist they don't for all subtypes of the flu, we can't be sure which strains will come each year, a pandemic is always a possibility. . For those who think otherwise, in 2009 H1N1, or the swine flu, spread rapidly throughout the world and killed between 151 and 575 thousand. A very likely candidate for a pandemic flu virus is the dreaded Avian flu, H5N1, this virus already infects a few thousand humans a year. However, what makes this virus so scary is that it is so easily spread by birds and is transmissible to humans with 60% fatality rate. Therefore, it is imperative to understand how the flu spreads and who is at risk for infection.

For our project we have tried to visualize a few graphics to show relevant information on Influenza. We have decided to show an "animation" of the flu spreading across the US, that way we can see the pattern of movement. We would also like to show which populations are more susceptible to infections. Then finally we would like to see if as the flu spreads to a greater portion of the population if it causes a higher symptom rate. By even answering a few of these questions we can hope to gain insight into the danger that is Influenza.

Related Work: Anything that inspired you, such as a paper, a web site, visualizations we discussed in class, etc.

One thing that inspired this project is the commonality of the influenza virus, and its implications. While humanity has managed to eradicate a number of viruses with vaccines, we cannot prevent the flu virus, even with frequent vaccinations. The quick speed at which the flu infects new populations and mutates, makes it a frequent threat to human life. The main graphic, the map, was inspired by a visualization of traffic data in a city, overtime you could watch the flow of traffic as it moved through the city, creating large jams then slowly spreading and dissipating. Ideally, this is how I pictured the map with flu infections starting in one city, growing and slowly spreading out along major traffic routes, before dying off completely within the initial location. It would show the spread from major city to major city and then from the large cities to the suburban areas before disappearing in the region. However, limitations in the number cities and states with data, will not make this possible to show.
One visualization inspired our graph number 4, showing the proportion of age groups that have influenza. The graphic is shown below:



However, since we did not have the deaths by influenza, we determined that we could still show the proportion of cases, to determine which populations were at higher risk for the flu, while simultaneously limiting ourselves to the more severe cases, because to be in our dataset the patient had to visit the hospital for the flu. However, to improve on this graphic we plan to overlay a line showing the proportion of the that each age group represents in the population.

<u>Questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?</u>

1. How does the flu spread across a country?
    a. How fast does it spread?
    b. How does it travel primarily?
    c. Who is the first to get infected?
    d. How well do vaccines prevent the spread of the strains they are intended for?

    e. How well do vaccines help prevent other strains?

2. What are the connections between the locations of certain birds known to transmit influenza, and known human influenza patients?
3. As more people are infected, are immunocompromised populations more susceptible to infections?
4. As more people are infected does the frequency of symptomatic cases increase as well?
5. Are there any noticeable connections between the spread of influenza and the characteristics of the spreading strains?
6. Do some strains cause more symptomatic infections than others?

These questions have evolved a lot over the course of this project. Initially, we started with much larger questions, and have since narrowed our scope. At the same time we have looked at our initial questions and the data and realized that we may need to new questions that can give insight to help answer the larger questions. For example questions 3 & 4,while inherently different, we cannot answer three with our data, but we can answer 4 and use our answer to hypothesize a response for 3. If we discover that yes, as the ratio of infected to uninfected is correlated to the ratio of symptomatic cases, then we can postulate that rise in symptomatic infections is because the virus is coming into contact with immunodeficient populations more readily. Overall, due to the amount of data, we are limited in the number and accuracy at which we can answer these questions.

Data: Source, scraping method, cleanup, etc.

We received our data from a multitude of sources. The primary data we started with was the Influenza Research Database found here:
https://www.fludb.org/brc/search_landing.spg?decorator=influenza

From this website we found datasets for both human and avian flu infections. The human data here is majority of what we are trying to work with. For the human and avian data we first filtered it to a specific date range, June 2012 - June 2013. This was one of the flu season in which we had the most data, approximately 1000 human samples.

The human data also had many extraneous categories so we limited it to State, Strain, Collection Date, Age, Gender, Temperature, Vaccination status, and symptoms. However, most of this data is in a form which we can’t work with. Therefore, we would need to edit, and create new values. First we took the collection date and made it into a numeric for the number of days since 1970, this way we can iterate through days, add days and compare dates. Then we added Age which we just made to convert from a string to a numeral. Then we utilized age and collection date to create what we considered to be the end of a contagious infection, 7 days for normal adults, and 14 days for immunodeficient groups. Then we added both vaccination status as a boolean, and temperature as a numeral. Then we had to search for particular symptoms in a symptoms array, find if they had a set of particular symptoms, fever, cough, and myalgia (muscle weakness), then convert that to a boolean for later use.

For the avian data we needed much less information, just collection date, state, and subtype. We used collection date to produce an end date for a contagious infection, and then we simply imported the other values as is.

There were a few other datasets we used for this project, to create the projection of the US, we used this Geo JSON file:

http://eric.clst.org/Stuff/USGeoJSON

This simply holds all of the information to produce a mapping of the US by Latitude and Longitude. This will be used with another dataset we found online, which is populations of every state, with Latitudes and Longitudes for the state capitals. However, to save coding work we manually altered this data to only have the states with Influenza cases and then we manually input it into the javascript. We also needed data on age groups and the percent of the overall population that they represent. We found this data through the US census and manually inputted it into the javascript. Lastly we needed information, on which strains/subtypes were predicted by the CDC for the 2012 flu season. This was manually inputted as two strings.

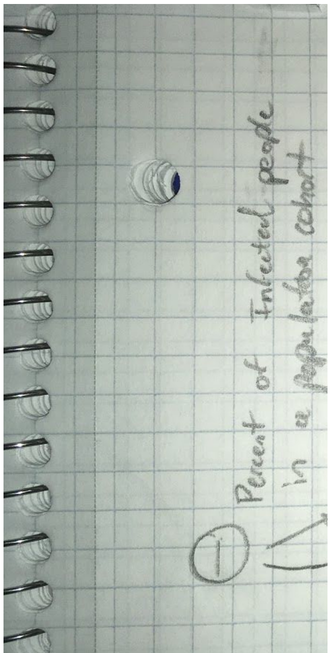
Exploratory Data Analysis: What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

Initially, for data analysis we performed queries on the data to aggregate it by state, then we produced bar graphs with the data to compare the quantities of different traits. From this data we saw a few different points of interest. First we noticed that we didn’t have data for every state, or city. Then we noticed that the amounts of data greatly varied between states. Then we noticed that all of the counts didn’t work because not all of the data is in the same form. Then we noticed that not all of the subtypes were common between states, for example, some states only had few subtypes not found in other states. Due to this information we began to realize that the scope of our questions and design might not be the most effective due to the lack of data. For example one of the biggest issues is that our map graphic was intended to show the spread of influenza across the US overtime, but a lack of equal quantity data from state to state may make it difficult to see a trend. Additionally, due to the amount we also realized that comparing the amount of infections with the area's population would no longer be effective because the relative sizes, a few thousand samples and a few hundred million people.

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

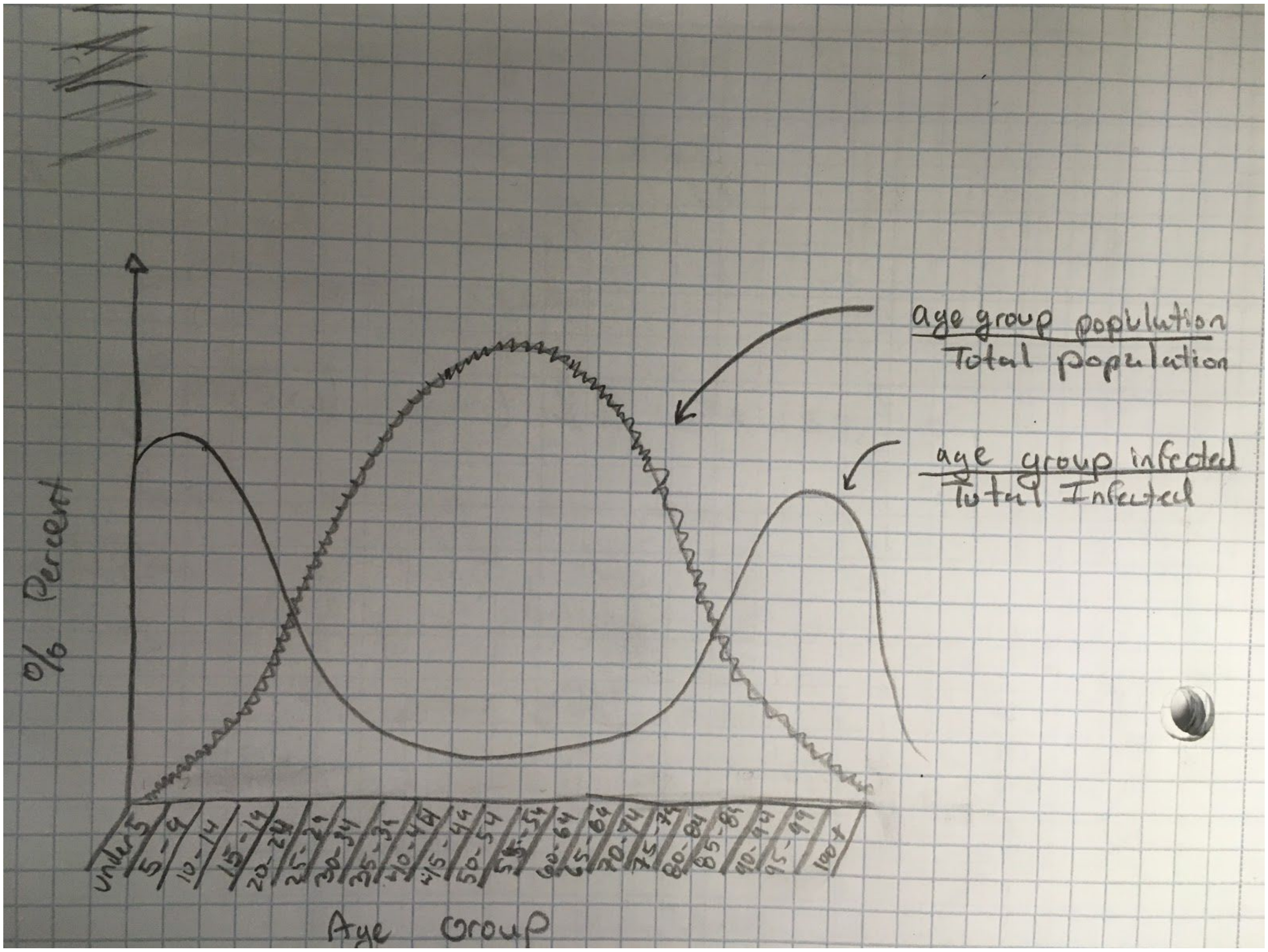
After getting our first visualization to work, graphic 2, displaying active human infections and active avian infections over time, we faced a difficult decision in our design. We found 3 distinct peaks, one in the center for human data, and two on the outside for avian data. However, what was truly confusing about this data is that the human data dropped off to zero infections, after and before either of the avian peaks ended or began. Therefore, we found that the avian data may be extraneous to show, when it has no noticeable interlap. However, after having looked at the graphic in many formats we decided to keep all the data and display all three peaks. We decided on this because the original intent of the graphic was to show a correlation between Avian flu infections and human flu infections. While what we found showed no correlation, in instead showed that the flu season for humans and the flu season for birds are in two separate times, of the year. It is noticeable that while human flu season takes place in the winter, avian flu season happens in the summer. Therefore, becoming infected with the flu is more about the time of year and season than it is about contact with animals or humans.

We had also decided to change the fourth graphic, shown below:



The goal of this graph would be to show what percentage of the total age population, that was infected with the flu. However, after looking at our dataset, we realized this comparison would be inaccurate and show essentially nothing. This was due to the dataset containing only about a thousand values, and the difficulty in finding the actual percent of the population that each age represented.

Therefore we decided we needed to change the graphic, first we changed the X-axis to be by age cohort, every 5 years. We added a line to show the percent of the total population that the specific age represents. Then we decided to change the existing curve to show the percentage of the infected in an age group out of the total infected. This way the trend we are trying to show, higher peaks toward older and younger populations, is not hidden by our sampling size. The new graphic is designed to look like this:



For the 5th graphic, the bar chart displaying the quantity of infections by strain we decided to also show the avian infections as well, as a red bar, clustered with human data by strain. This way we could directly see which strains infected birds humans and both.

Lastly, as we have been working on our project, we have considered our first design of placing the graphics and information onto multiple pages of our website. However, this would force the user to click through pages to find relevant information, and it would end in a cluttered graphics page with no way for the user to discern useful information. Therefore, we decided to implement a simplistic version of a scrolly-story, and have everything on page and and as you scroll down you are lead through of description of the data and the Influenza virus. This will hopefully make it easier to interpret and to gain insight from for all levels of user. To help accomplish this goal we are going to use bootstrap template. This will help to make our design more visually appealing and help tell the story of the how Influenza spreads.

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

We want to be able to filter the majority of the visualizations by a few different factors, Age group, strain subtype, gender, state, and symptomatic cases. This will hopefully allow us to narrow the data and see trends not represented in looking at the entire data sample. This will be done by selecting values in the other graphics. However, not all of the visualizations will be able to be filtered by every option. As we look through the data quantities and produce the respective graphics we are seeing a conflict between what we would like to show and what can be shown. A perfect example of this is the second graphic, showing a comparison of human and avian infections over time, initially, we would have liked to filter this by all options, however, because the avian data does not have gender this is not possible. Otherwise all of the graphics will be influenced by any selection in any graphic.

Eventually, as we complete all of the graphics we will fill this in with images, respective filtering options, and explanations as in the proposal.

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

We will answer this question with images and data once the website is completed.