CS 171 Process Book
Project Name: "Healthcare, Meet Data.gov"
Members: Jake Dorabialski, Daniel Jung Dong,Wentao Xu

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

There are many issues with the US healthcare system. Healthcare costs have been rising faster than inflation over the past fifty years, and spending now account for a fifth of US GDP. Despite this, the World Health Organization ranked the US 37th in quality of medical services. To help reign in on wasteful spending and to increase overall transparency, the US government has established an open data initiative to make to make hospital spending and performance easily accessible to the public. As a result, we decided to look into the publicly released data on www.data.gov to explore the relationship between hospital spending and the effectiveness of healthcare. By creating a data visualization tool based on these data, we hope to help anyone looking for cost-effective healthcare to contrast and compare hospitals on the local, state, and national level. We hope that in addition to helping patients make a better choice when selecting hospitals, we want to show trends highlighting the factors associated with the poor performance of some hospitals, and so reveal patterns in the data that could be important for policymakers.

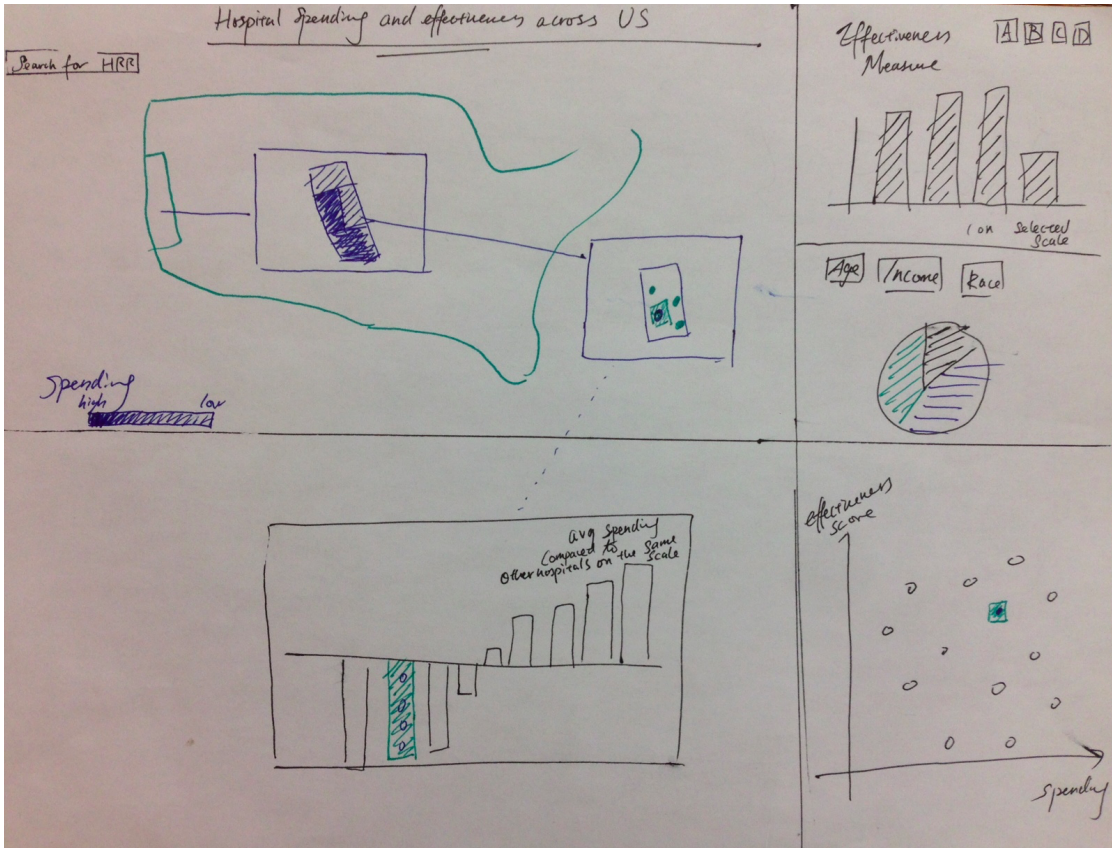
Project Objectives. Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits.

Hospital spending and healthcare effectiveness vary greatly across different geographic regions. This geographic variation is affected by demographics, socioeconomic status and government policy, to name a few. Through our project, we hope to first shed light on the differences in the effectiveness and cost of healthcare, which will be immediately helpful in allowing potential consumers compare and "shop" various hospitals in their

region. Specifically, we hope to answer the following questions through our visualization tool:

- Which hospitals are high-performing? Which ones are low-performing?
- Which hospital within the selected county is the best hospital to go to?
- How does the selected hospital compare to all the other hospitals in the same county, state, or country?
- How does the spending and effectiveness differ among different types of hospitals?
- What are the similarities between high-performing hospitals? And what are the similarities between low-performing hospitals? What can the low-performing hospitals learn from high-performing ones?

Implementation Strategy: We plan to create an interactive map of the US where users can zoom in and out of more disaggregated district levels (i.e., from state to county level). Users can select from various measures of healthcare effectiveness and contrast it to cost at various levels of regional aggregation. We plan to make this clear through a heat map visualization. Lastly, at the most disaggregated level we intend to show individual hospitals and their relevant metrics.

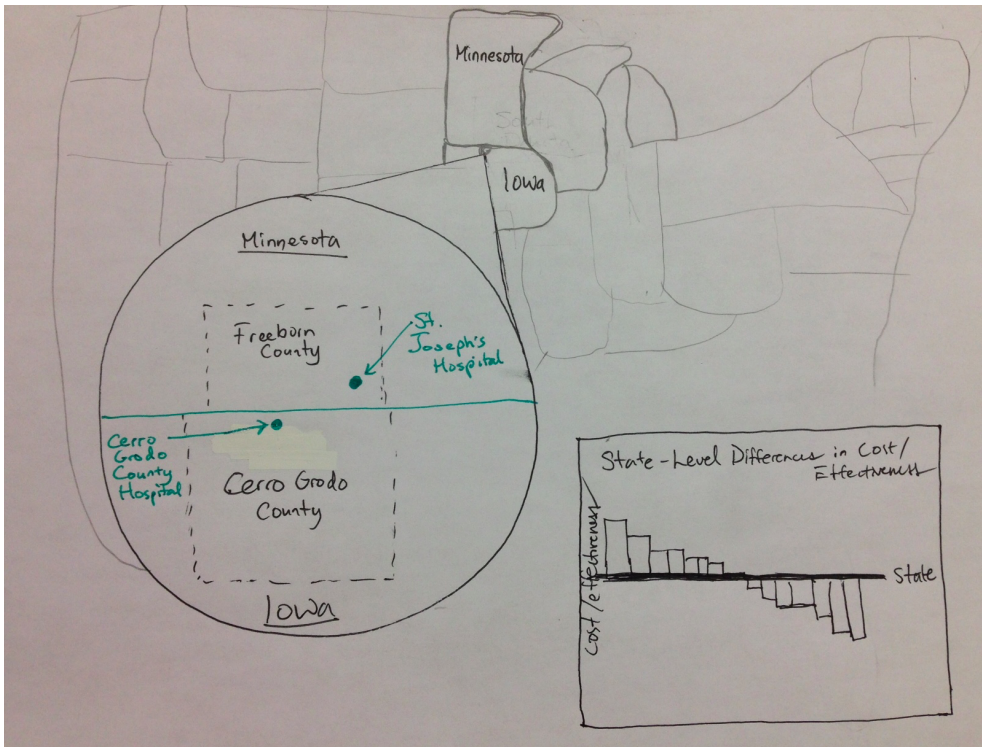


In the second step of the project, we plan to incorporate different visualization techniques to show which factors are associated with healthcare effectiveness with a focus on three factors -- **demographics, socioeconomic status, government policy**.

- How do demographics (population distribution by race, age) correspond to hospital spending of a district?
- How does a regions income level correspond to the effectiveness of healthcare?
- To what extent is hospital effectiveness associated with the price of healthcare? Is good healthcare solely determined by its cost?
- Are there some factors in districts associated with high hospital cost-effectiveness?
- To what extent does government policy help improve (or reduce) the quality of health care?

Implementation Strategy: We intend to layer the initial heat map of the US with the different statistics on demographics and income levels. Users can navigate between the different statistics through a switch panel. We also plan to supplement the heat map for some of the variables such as income, with a side panel showing the wage inequality at a specific district level (through a simple histogram).

In analyzing the effect of government policy, we intend to perform the following case analysis. To better isolate a causal effect of government policy, we plan to look at effectiveness outcomes for hospitals with similar characteristics of the population they serve, but under different set of government regimes due to the region they serve. By looking at pairs of hospitals, separated by a state border we intend to again present effectiveness measures, but how they are affected by government policy rather than demographic and socioeconomic determinants.



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

A. Demographic Data
    a. Description: The database contains 2010 county level data on population distribution by age, race, and poverty level.
    b. Source:http://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer
B. Income Distribution
    a. Description: The database contains the distribution of incomes at the county level. It also provides data on unemployment and number of people on welfare.
    b. Source: http://www.irs.gov/uac/SOI-Tax-Stats-County-Data-2012
C. Hospital Spending Data
    a. Description: The database contains the Medicare Hospital Spending by each care episode.
    b. Source:https://data.medicare.gov/Hospital-Compare/Medicare-Hospital-Spending-by-Claim/nrth-mfg3
D. Hospital Type
    a. Description: The database contains hospital type information (Government, Non-profit, Proprietary) for about 4800 Medicare certified hospitals.
    b. Source: https://data.medicare.gov/data/hospital-compare
E. US Healthcare Spending
    a. Description: The database contains the breakdown of US healthcare spending from 1960 to 2013.
    b. Source:http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html

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?

A. Demographic Data
    a. The database contains information about age and race distribution in columns in which each row represents a county. Thus, we would need to iterate through each row and store the distribution data in a dictionary where the key is the county name and value is an array of dictionaries.
B. Income Distribution
    a. This will involve extensive data processing as the income data is separated into different files for every state. Furthermore, within each file,

the income data is not structured in a way that makes it easily accessible (no consistent index values, except for a country FIPS code.

C. Hospital Spending
    a. The database contains the total hospital spending for each episode of care. Therefore, we would need to iterate through each row and store the information about the hospital spending in a dictionary where the key is the hospital name and value is the spending amount.

D. Hospital Type
    a. The database has about 4800 rows with each row representing different hospital. Thus, we would need to iterate through each row and store the information about the hospital type in a dictionary where the key is the hospital name and value is the type.

E. US Healthcare Spending
    a. The database contains annual spending for each year from 1960 to 2013 in columns. Thus, we would need to iterate through each column and store the spending information in a dictionary where they key is the year and value is the spending.

Visualization. How will you display your data? Provide some general ideas that you have for the visualization design. Include sketches of your design.

Our design will include a central “heat map” of the United States with a color-variant view of average hospital expenditure across different different states and counties. Users will have a click-to-zoom option on specific state and county. This option will create a new layer of the zoomed-in heat map on top of the original. On the sides there will be bar charts, area charts to supplement information on effectiveness of treatment, income, age, and race distribution, and the hospital’s spending as compared to hospitals on state, country, or individual level .

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

The project must feature a way to magnify certain state or district selected from the heat map when clicked on. Furthermore, the project must be able to feature bar and area charts on the side and update itself dynamically according to the scale selected from the central heat map (from federal to state and to county level). It is essential that we represent the different demographic and socioeconomic variables in the main US map, as well as peripheral charts (i.e., various income distributions). Another core feature is having an interactive visualization that shows the differences in effectiveness and cost that arise due to government policy differences. This visualization will be supplanted with another chart showing the the average absolute differences between states. We may correlate this with measures of state-level bureaucracy (or health care restrictiveness) just to validate that it is in fact government policy driving these differences.

Optional Features. Those features which you consider would be nice to have, but not critical.

One of the optional features is looking at the effect of government institutional structures on healthcare cost and effectiveness. One of the optional features around this would involve creating a variance plot of which state pairs have the widest gaps.

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.

Weeks	Tasks
4/4-4/10	● Process the data. ● Learn about how we can load the data onto a map like how it was done in “Century of Corn” ● Load the data onto a map given the city and state.

4/11-4/17 **(1st Milestone Deadline)**	● Make sure click-to-zoom works to allow zooming onto state level and then onto county level. ● Assign color coding for different types of hospitals
4/18-4/24	● Create visualization for Effectiveness, Demographic, and Comparison. ● Make event handler work for interactive display across different views
4/25-5/1	● Implement any optional features ● Do the write up ● Prepare for presentation
5/2-5/4 **(Final Project Due)**	● Finish any last minute adjustment