Investigating Bicycle Fatality Patterns

Saad Ismail, John Nakano, Hannah Kim, Siddharth Raja

Document Outline

1. Project Motivation / Purpose

Biking is being taken up by an increasing number of Americans every year not just for leisure or exercise but also as a substitute for traditional means of commuting to work. With biking gaining importance as a serious mode of commuting, it is important to ensure that it be a safe option for people. Every year, close to 1000 bicyclists are killed on the road throughout the US. With the number of bike riders increasing, these numbers are also likely to rise unless we identify the major causes and circumstances and take preventive steps. So, it is necessary to get some sense of what makes biking unsafe. In this project, this is what we attempt to understand through our data and our visualization of it.

To better understand these circumstances, we analyze a dataset of bike accidents across the US from 2008-2012. We consider time of day (which gives us an indication of lighting on the road), weather conditions, what part of the road the bike-rider was on (example: bike lane, normal road, crosswalks, etc.) and finally whether the state in which the accident occurred allowed bikes to be ridden on sidewalks. The state-wise sidewalk laws are interesting because bikers often keep switching between roads and sidewalks while riding their bikes, unless the state law disallows it. It would be fascinating to see how this behavior (indirectly influenced by sidewalk law) affects the number of crashes. We may also be able to tell if riding on the sidewalk is more dangerous compared to on road.

Most visualizations related to road fatalities are either dedicated to motor vehicles or pedestrians with bike riders being largely ignored. Furthermore, such visualizations generally revolve around geographic locations to reveal best or worst location by number of fatal crashes. Several works plot accidents on map-based visualizations.[1,2,3,4] While these works effectively show geospatial hotspots, we cannot observe the relations between fatalities and other risk factors with their visualizations. Others have tackled bicyclist trends across the US as they commute to work or have looked at bicycle sharing data within their cities.[5,6,7] However, these works either only look at how few data points influence a single demographic, overlook pertinent contextual information that might provide a deeper understanding of bike-related

[1] Mapping Car Crashes in the UK, http://www.mapsdata.co.uk/portfolio-items/traffic-accidents-uk/
[2] NYC Crashmapper, http://nyc.crashmapper.com/
[3] BBC Every death on every road in Great Britain 1999-2010 http://www.bbc.com/news/uk-15975720
[4] UK Road Accident Map, http://www.osbornes.net/media/uk-road-accident-map/
[5] Bicycle Commuting Trends in the United States, http://envs.uoregon.edu/bicycle-commuting-trends/
[6] Citi Bike Data Visualized, http://linepointpath.com/111242/2771111/work/citi-bike-visualization
[7] Bicing Spider: bicycle sharing data visualization (Beta), http://blog.jpcarrascal.com/2011/03/bicing-spider-bicycle-sharing-data-visualization-beta/

issues. In addition, to the best of our knowledge, none of the existing visualizations investigate the relation between vehicle accidents and safety laws.

Our visualization would hopefully answer some of these questions and other similar ones:

- How do sidewalk laws affect fatality rates of bicyclists across the states in US?
- Are there specific times of day where it is most dangerous to ride bicycles?
- What kind of weather should bikers avoid. Or in bad weather, would it be better to ride the bike on a specific section of the road?
- Do states that prohibit riding on sidewalks have more or less accidents on sidewalks? What other locations do the accidents occur in?

Note that several machine learning (ML) techniques can be applied to solve these questions. For instance, we can use various modeling techniques to predict how safe it would be to ride a bicycle in given conditions. However, a visualization leverages human perception and can easily find multiple patterns while most ML techniques aim to find a global pattern. Also, to investigate the relations between fatalities and risk factors using ML, a user must specify which risk factors or combinations of risk factors are of interest, which is nearly impossible without prior knowledge of data or huge computation capabilities. In addition, in case of outliers (e.g., higher/lower fatality numbers than expected), visualization allows more in-depth analysis through interactions but most ML techniques does not offer much information other than it being an outlier.

2. Users & Tasks

Our visualization targets bike enthusiasts and policy makers. Bike enthusiasts ride bikes frequently for commuting and would be interested in the safety of biking in relation to external conditions, time of day, and locations. Policy makers consist of people in state and federal governments. They are interested in how state sidewalk laws and locations on road affect bicycle fatalities.

Bicycle enthusiasts will use the visualization to compare categorical variables (location of crash and weather) to determine any relationships or patterns. They can determine what combination of these variable types is the best or the worst for bike riding. This data is also split up state-wise and by time of day. So for example a biker in Georgia can see what the fatality numbers are for riding in rainy weather on an intersection with a crosswalk. If that user wants to compare this with the numbers from California, they can do that too.

Policy makers will also be able to use the visualization to determine relationships between categories. However, they will find “Law Mode” the most relevant to them. “Law Mode” will allow the policy makers to visualize how laws affect fatality rates. Our visualization allows the user to turn on “Law Mode” which will affect the heat map overview and the detailed line graphs. Policy makers will be able to see how categorical variables and laws together affect bicycle fatality rates. In addition, the line graphs can be used to compare states will different sidewalk laws. Therefore, if a policy maker is looking to change sidewalk laws for one state, he or she will be able to compare how other states with various sidewalk laws are doing. Policy makers will also be able to see how where most crashes take place (bike lanes, sidewalks, and etc). Depending on this data, they can enact new laws, implement bike lanes, and etc.

In addition, activists who are dedicated to promoting safe biking across US, can also benefit from this visualization. They can use this visualization to determine which laws affect bike safety to influence lawmakers and the general public accordingly.

3. Dataset & Attributes

Due to the amount of attributes and categories we are visualizing, we require a rich data set from multiple sources. We will be using the FARS (Fatality Analysis Reporting System) data set, BikeLeague data, and census information for this project.

Fatality Analysis Reporting System (FARS)

FARS[8] is a nationwide census maintained by the NHTSA (National Highway Traffic Safety Administration) that provides data on all fatal traffic crashes in the US. The FARS dataset contains fatality reports (of motorists and non motorists) with various attributes for all states. The dataset spans from 1994 - 2012 however we will be focusing on 2008 to 2012 due to the availability of state laws data. There are about 3300 fatalities for the years 2008 to 2012. We will be taking a look at data attributes related to number of bicycle fatalities, external conditions (weather), time of day, and location of crash on the road. These attributes are described in greater detail below:

Fatality Numbers

We will filter the number of fatalities by state and motorist type. These numbers will have to be normalized based on state population (explained in greater detail below).

Timestamp of Each Accident

The timestamp of the accident is present in raw form by date, hour, and minute. We will be leveraging it to split up the data set by 3-hour time blocks (12-3 am, 3-6 am, and so on). Time blocks we are condensing the amount of attributes there are on the visualization and allowing the user to easily visualize important times. Such as visualizing commuting/rush hour times are from 6am to 9am and 3pm to 6pm).

State that the Accident Took Place In

[8] "Fatality Analysis Reporting System (FARS) | National ..." 2010. 22 Oct. 2014 <http://www.nhtsa.gov/FARS>

We are given the state of which the accident occurred in. We will be using this data to allow the user to filter out accidents by states.

Atmospheric Conditions

Atmospheric Conditions consist of ten different conditions (clear, cloudy, rain, sleet, snow, blowing snow, fog/smog/smoke, severe crosswinds, blowing sand, and other).

Location at Time of Crash

This category details as to where the bicyclist was at the time of the crash. These categories (total of 15) consist of crosswalk (marked, unmarked), bicycle lane, sidewalk, shoulder, median, shared path/trail and etc.

The League of American Bicyclists

The League of American Bicyclists[9] have compiled bicycle state by each state for a research project for 2012. We will be using this data set to determine whether it is legal to ride on the sidewalk or not for all 50 states.

Sidewalk Riding Laws

These laws are not strictly defined as yes or no. Some states have laws have age restrictions and some states do not have sidewalk riding laws at all. For states that have age restrictions, most of them state that children are allowed to ride on the sidewalk (17 and under). We will be looking at laws as a strict yes or no. For states that have age restrictions, we will be determining if its legal to ride on the sidewalk as an adult. For states that do not have laws, it can be assumed that sidewalk riding is legal.

As mentioned earlier, the laws data set was for laws in 2012. To be able to visualize fatalities across four years accurately, we will need laws for 2008 to 2012. To solve this issue, we used information from the NCSL (National Council of State Legislatures)[10] to determine if any laws have changed anytime between 2008-2012. It turns out that no

[9] "State Bike Laws | League of American Bicyclists." 2013. 22 Oct. 2014 <http://bikeleague.org/content/state-bike-laws-0>

[10] "State Traffic and Auto Safety Legislation Data and Information." 2013. 22 Oct. 2014 <http://www.ncsl.org/research/transportation/state-traffic-safety-legislation-database.aspx>

new bicycle related sidewalk laws have been enacted between 2008 to 2012. Therefore, we can safely assume that the bicycle laws for 2012 are valid from 2008 to 2012.

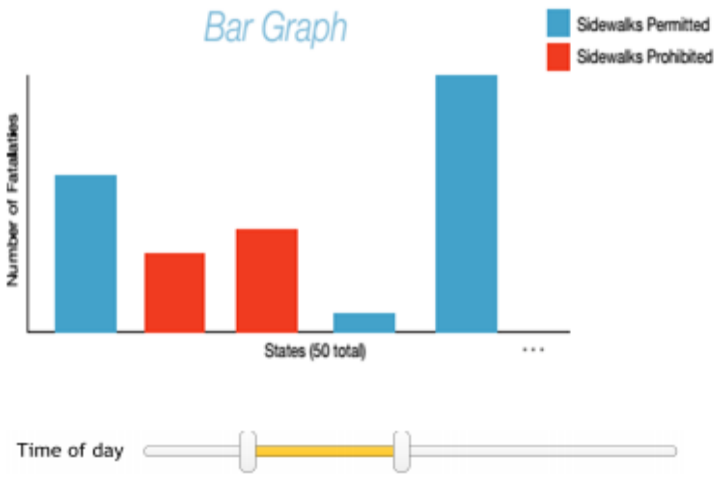
US Census

The US Census population data will be used to normalize the fatality numbers for each state based on the state population. We will be interpolating and extrapolating population data for 2008 - 2012. To normalize, the fatality numbers will be represented as ‘incidents per 500,000 people’. We choose 500,000 because the state of Wyoming with lowest population has roughly over 500,000 people (2012 estimates). This would give us a sense of comparison between say California (38 million population) and Vermont (600,000 people).

4. Designs That We Considered

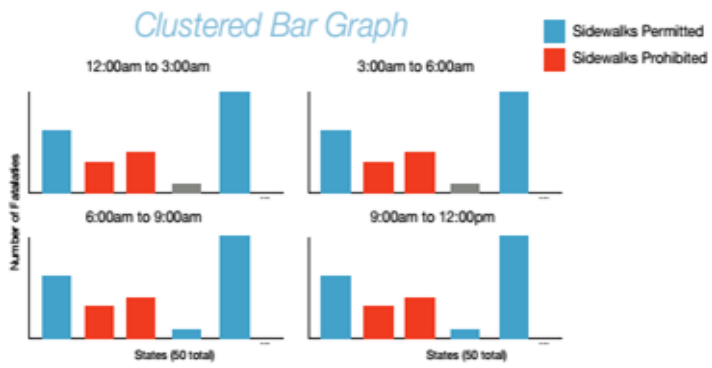
The designs we considered and presented in our poster session span across multiple visualization types (bar graphs, heat maps, scatter plots, calendar heat maps, and etc). A description and a brief analysis about each visualization is included below.

Bar Graphs



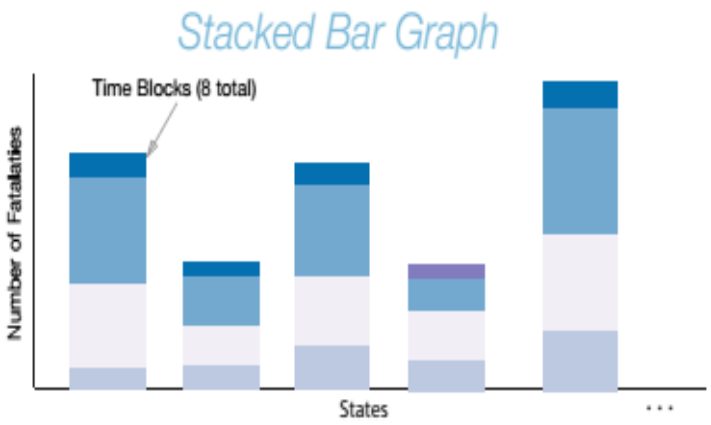
The bar graph shows the number of fatalities of each state. A time slider allows filtering by a specific timeframe. Bars are color-coded with the sidewalk laws of all states for easier comparison between different laws. The x-axis can be ordered by the number of fatalities or alphabetically.

Major Drawbacks:
With 50 states, the graph is too long. Would only serve as a good overview of the data. Not useful to visualize data by multiple time slots simultaneously.



The clustered bar graphs show the number of fatalities of each state for each time grouping. Better than simple bar graphs to analyze trend across time blocks. Bars are color-coded with the sidewalk laws of all states for easier comparison between different laws.

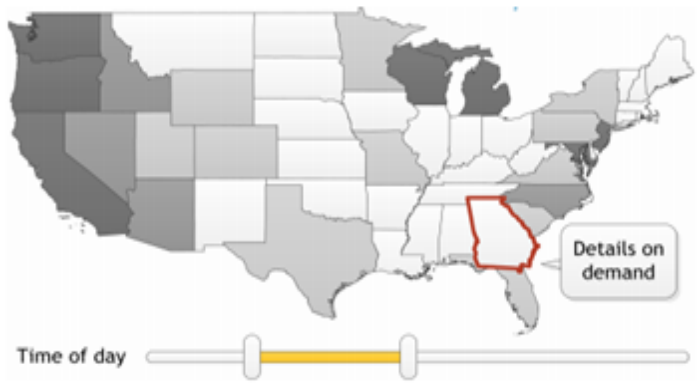
Major Drawbacks:
With 50 states, the graph is too long. Since they are already split by time slots, they can only visualize one combination of factors at a time. Adding more factors would greatly increase the numbers of clustered graphs required causing cluttering of the screen.



The stacked bar graph shows the number of fatalities of each state, stacked by time groupings. Bars are color-coded with the time blocks. The x-axis can be ordered by sidewalk laws or alphabetically.

Major Drawbacks:
It is hard to see fatality trends between different laws, and a comparison between time blocks is not easy.

US Heat Map



Colored Map

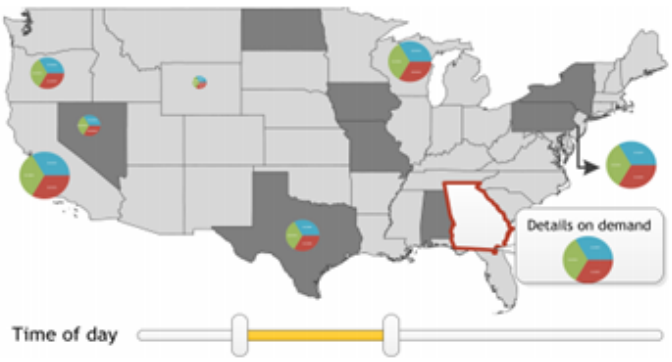
The colored map uses shading to represent the number of fatalities to see geospatial trends. It also has a time slider to update the shading to represent data for the selected time period. A user can click or hover on a state for more info. This variant of the colored map allows for easier recognitions of state by shape and geospatial relations.

Major Drawbacks:
It's difficult to recognize the smaller sized states. It's also difficult encode both state laws and fatality numbers on a map.

Colored Map (with Bars)

This variant of the colored map would use shading to encode state sidewalk laws allowing for easy determination of the laws by state. States would have a bar graph or similar encoding to represent number of fatalities.

Major Drawbacks:
It is difficult to compare fatality numbers between states with the different bars due to the different sizes of the bars as well as the different baseline of each bar. The bars also have an undesirable 3D effect.



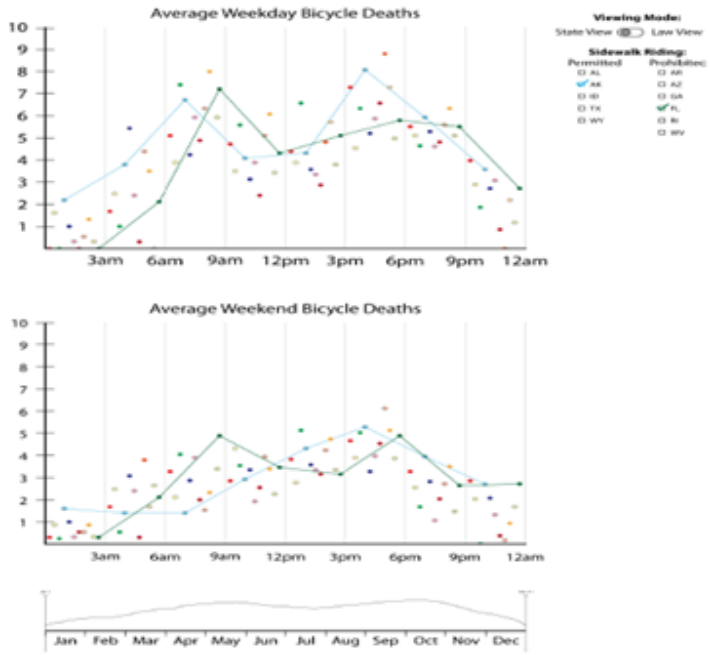
Colored Map (with Pies)

The colored map with pies uses shading to represent sidewalk laws just like the previous option. State would have a pie chart to show the number of fatalities between weekday and weekends. It also has a time slider to update the shading to represent data for the selected time period.

Major Drawbacks:

It is difficult to compare fatality numbers between states with the different pie charts due to the different sizes of each one. Also pie charts are more effective if percentage values are used with the total being 100%. Using pie charts to depict numbers is not the best option.

Scatterplot combined with Line Graphs

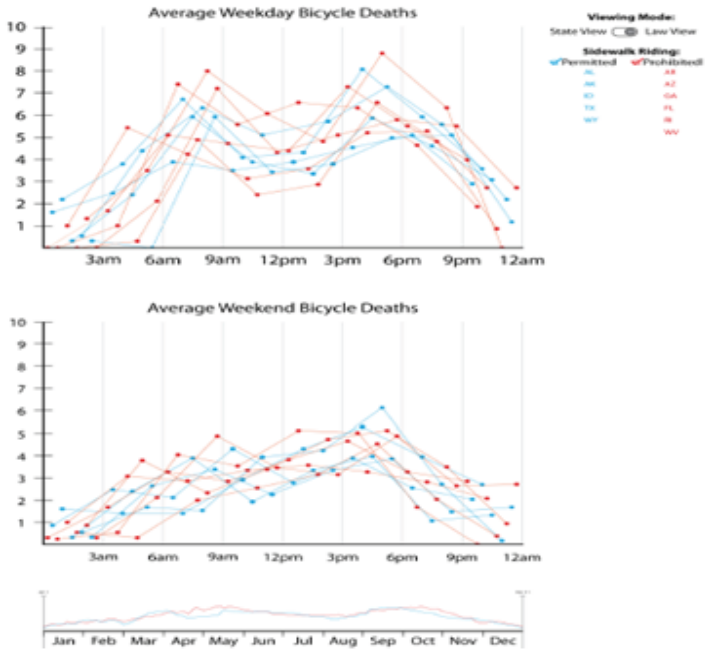


Selection based on individual state ('State View')

The scatter-plot would be useful for displaying average fatalities for weekdays and weekends for the selected time period. States are encoded with different colored dots. Data points have shape-encoding to identify their position on sidewalk riding depending on the state they represent.
This is good for displaying general trends for bicycle fatalities across the year in the overall view.
Visualization is separated off into categories (states vs laws) so user can easily switch between the two modes.
The picture on the left shows a typical state view.

Major Drawbacks:

The color encoding would require 50 different colors for dots which would lead to visual clutter. Adding that with shape encoding for each state would also further complicate the problem. Dots also suffer from overlapping problem.



Selection based on sidewalk laws ('Law View')

The picture on the left shows a typical law view. The states are colored by whether the sidewalk laws allow bike to be ridden on the sidewalk.
The color encoding problem is solved by choosing only two colors and also getting rid of the shape encoding.

Like the previous illustration, we can remove or add states from the view as per choice. The timescale at the bottom allows the user to select any one time frame.

Major Drawbacks:
The overlapping problem is still not solved. However, in our final design we do incorporate the line graph and get rid of the scatterplot. We only plot the line for all states that have been selected.

5. Feedback from Poster Session

A major feedback we received from our poster session was to try and utilize more attributes from our data set to probably expand our current visualization scope. This made us think about choosing location of the biker at the time of crash and atmospheric conditions, which we had not considered before. With these additional attributes we needed to find a clean way of integrating them with the original visualization or creating a new sub visualization. We ended up on creating a heat map overview (of categorical variables) which we are now using to depict these attributes. We now have a richer visualization which can hopefully help the user answer more insightful questions with those parameters.

Another feedback was to focus more on the interactions of the final design. We have therefore given much thought to the various components in our visualization and also their interactions through which the user would accomplish tasks. We have been able to come up with specific scenarios for user tasks and also refine some of the questions that the visualization will answer with these interactions.

Defining this interactions and incorporating new attributes into the system also makes it unique to address this dataset related to bike accidents and some of the questions we could answer from it. Current InfoVis toolkits such as Tableau could not provide us with such detailed interactive visualizations where we depict factors across a category heat map and show state-wise trends for various combination of factors on a line graph. Furthermore, both the heat map and line graphs are dependent on a time period selected by user. Clearly, such high level of interactivity and filtering that addresses this specific dataset cannot be achieved by existing toolkits.

6. Design Choice and Intent

Our overall design will be set up in one browser window and in vertical orientation. In this window, there will be two parts: an overview & controls view (Fig. 1(A)) and a detailed view (Fig. 1(B)).

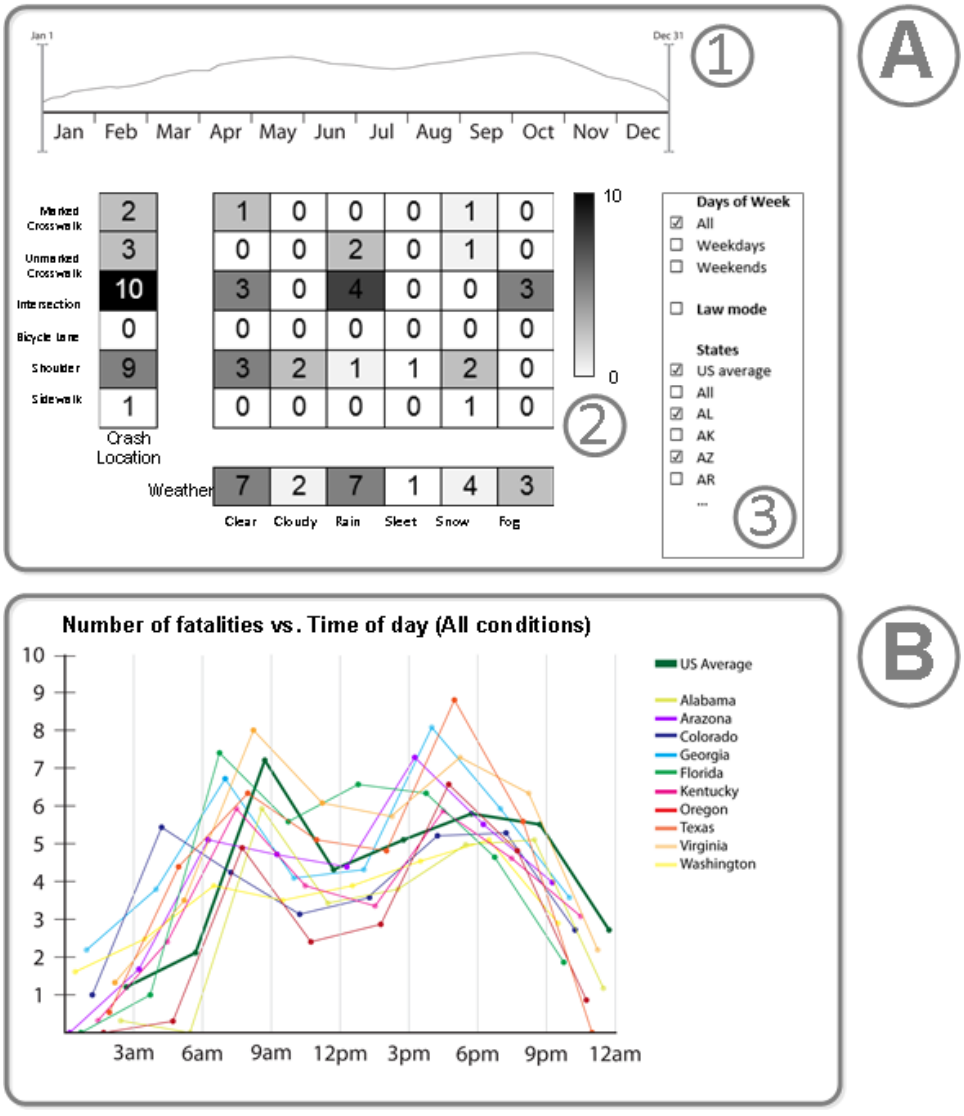


Fig 1. An overview of our design. (A) the overview and controls view; (B) the detail view. The overview and controls view have a time slider with an over-layed line graph showing the overview of fatality trends over year, a heat map overview/filter to show fatality patterns by weather conditions and crash locations, and a control box that provides various filtering options. The categories (weather condition and crash locations) are defined to the left and below the axes. The detail view displays multiple line graphs (fatality versus time of day) for filtered data.

Overview and Controls

This part will have two overview graphs and a control box. Across the top is a line graph with the total number of deaths on the y axis and the days and months on the x axis (Fig. 1(A-1)). This line graph also acts as a time filter over year. The user will be able to drag two sliders to select/filter a range of months/dates for more in-depth analysis. This year overview allows the user to identify high fatality periods and filter out the periods of lesser interest. This selection/filtering will affect the subsequent visualizations.

Fig. 1(A-2) shows a grid heat map overview (by categorical variables) that visualizes the number of fatalities happened in specific weather conditions and crash locations relative to the street (e.g. bike lane, intersection, sidewalk,...etc) as both shades and numbers on them. The y axis represent weather conditions and the x axis represent crash locations. The axes themselves are also heat maps representing the aggregated number of fatalities along the axes. That is, the y axis is shaded based on the number of fatalities in specific weather condition. Similarly the x axis is shaded base on the number of fatalities in a specific crash location. Additionally, this heat map has a filtering capability. When a user clicks a rectangle representing a specific weather condition or crash location, the scope of data visualized in the subsequent visualizations in the detail view will change accordingly. If more than one rectangles are clicked, additional visualizations representing the clicked portion of the data will be appended in the detail view.

Beneath the overview there will be a set of controls that further filters the data (Fig. 1(A-3)). These controls are for limiting the data to all days (default), weekdays, or weekends. Additionally, the user will be able to select the US average or select individual states for comparison against each other and against the US average. Furthermore, the “law mode” button gives an option to split the visualizations into two colors to color-code the lines by sidewalk laws (i.e., not permitted as red and permitted as blue) in the heat map (Fig. 2) as well as in the detail view (Fig. 4).

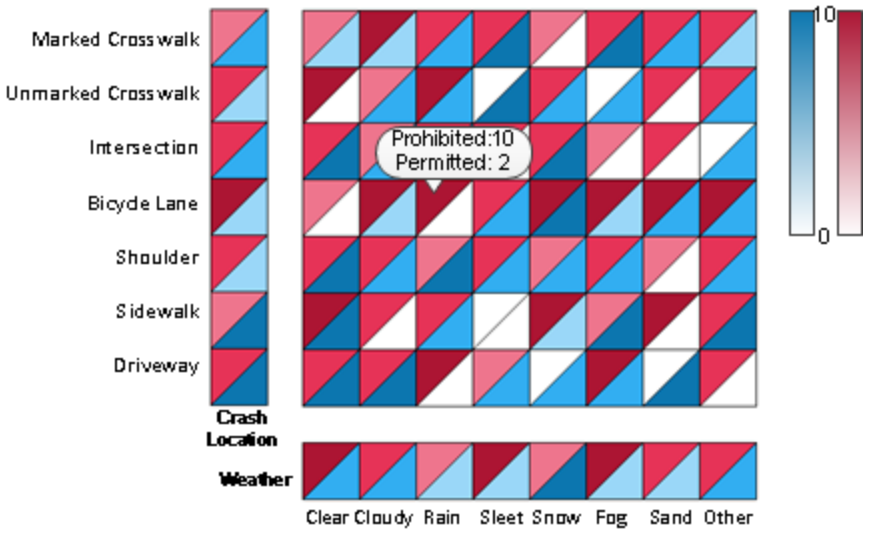


Fig 2. An updated heat map in the overview and controls view (Fig. 1(A)) when the law mode is selected in the control box (Fig. 1(A-3)). The intensity of blue and red shades in each rectangle represent the numbers of fatalities in its specific weather condition and crash location for the states that allow riding in the sidewalks and the states that prohibit it, respectively. A pop-up window showing the number of fatalities for both cases appears on mouseover. This picture is a mockup of the final heat map.

Detail Across Time of Day

The detailed view will show a line graph with the x axis segmented into time of day blocks whose initial time blocks are of 3 hour increments, which allows for easily separating critical time blocks of the day such as dawn, morning rush hours, mid morning, lunch time, mid afternoon, evening rush hours, dusk, and late night. The y axis will be the average number of bicyclist fatalities. The graph will have an option to show the US average across the time of day. For easier comparison, when a user clicks or hovers over a line, the line is highlighted as the other lines except the US average line will be grayed out. If there is two or more graphs in the detail view, states that are grayed out in the clicked graph are also grayed out in the other line graphs (i.e., brushing and linking).

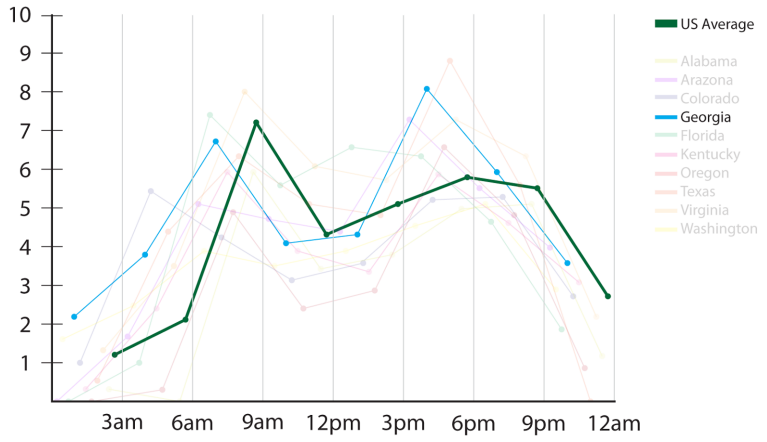


Fig 3. An updated line graph in the detail view (Fig. 1(B)) when a line is clicked or hovered over.

If the law mode is selected, it will show two average lines: one for the states that prohibit sidewalk riding, and one for state that allow sidewalk riding (Fig 4). Furthermore, in law mode, all individual state lines graphed will be one of two colors to easily associate and compare a specific state to the corresponding national average of all states that have the same sidewalk riding law. The specific average that is shown will depend on which condition box is selected in the category grid heat map visualization. If more than one condition box is selected, a separate line graph for each condition will be made. When a user has selected specific states in the filter controls, the states' average will be shown on the map as well. By having separate graphs based on conditions, it allows for easily comparing multiple states in a variety of conditions.

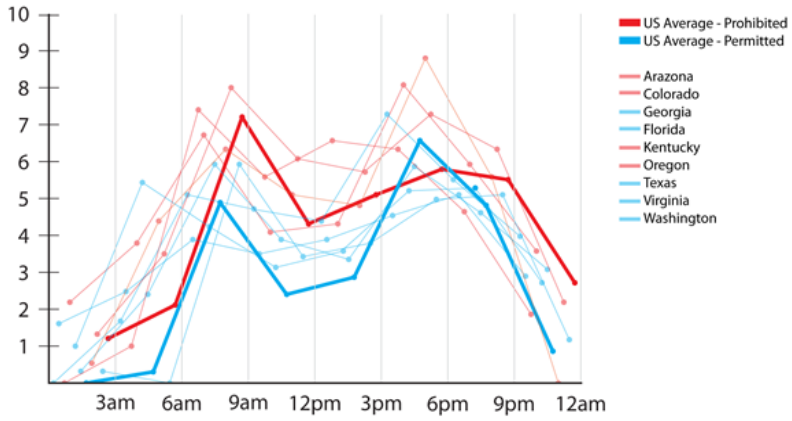


Fig 4. An updated line graph in the detail view (Fig. #(B)) when the law mode is selected in the control box (Fig. 1(A-3)). Lines for the states that allow riding in the sidewalks and lines for the states that prohibit it will be colored as blue and red, respectively. Instead of one US average line in non-law node, the updated graph will show two average lines for two groups of states.

Usage Scenario Example

Assume that a user wants to see fatality patterns during April and March. First, the user filters time using the time slider (Fig. 5(A-1)) and selects states in the control box (Fig. 5(A-3)). The heat map in Fig. 5(A-2) is updated using the filtered data. The user selects two rectangles of interests, which updates the pre-existing graph (Fig. 5(B-1)) and adds one more graph (Fig. 5(B-2)) in the detail view. Now the user can click or hover over multiple lines of interests for easier comparison. Whenever a line is highlighted in a graph in the detail view, the corresponding line in the other graphs in the detail view is also highlighted.

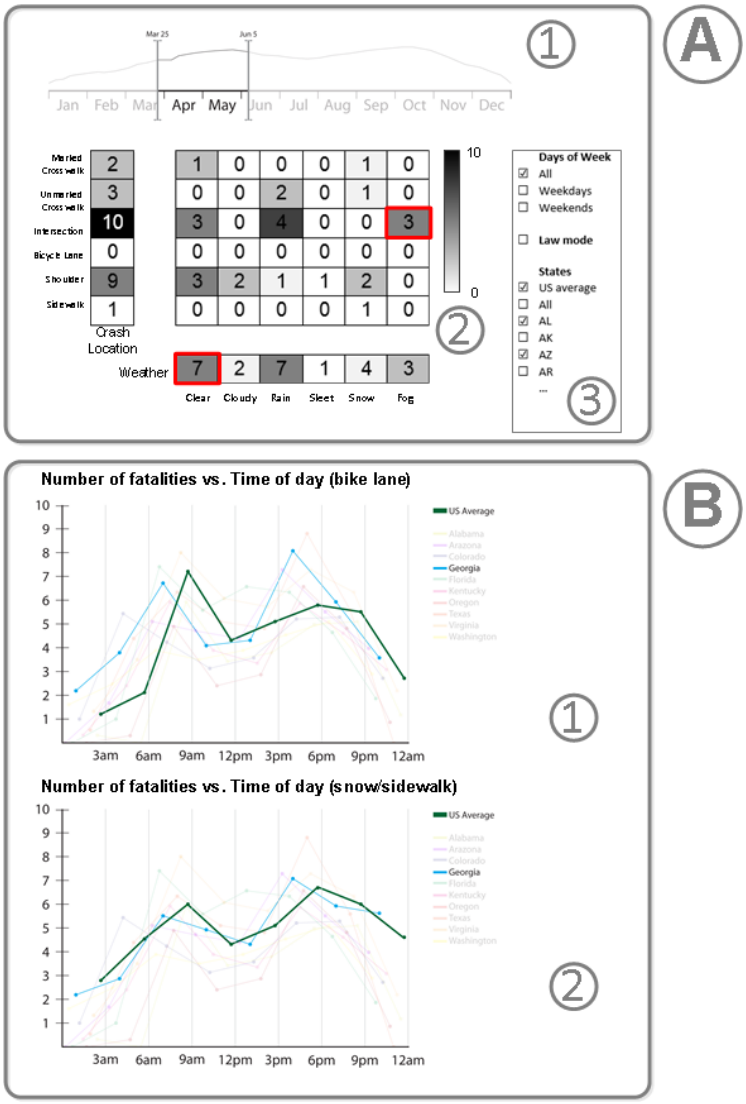


Fig 5. An overview of our design when a line is selected in the detail view when (A-1) a time period is selected and (A-2) two rectangles in the heat map overview are selected to show filtered data in the two combinations of categories in two line graphs (B-1,2).

Technical Details

D3 and Javascript will be used to build this visualization and its interactions. Since this requires rich interactions, we will be using a single browser window approach. This will make it easier for the user to understand how certain filters are affecting the multiple “sub-visualizations”. The dataset will be stored in memory and processed locally with D3. However depending on our findings with performance, we may choose to switch to a server and MySQL database. The queries will be processed on the server and stored locally. We will also be doing some pre-processing on the data set. This pre-processing includes compiling the state laws and linking them with the fatality data, filtering out any irrelevant information to make the dataset size smaller, and normalizing the data set beforehand.

Our visualization is split up into three different sub visualizations. The first one incorporates a line graph overview that spans across months in a year. This line graph will contain a selector that will allow you to filter through the different months. To build this we will be using D3 Line Chart APIs[11] to create the overview line chart and implement a custom solution for the selectors.

The second visualization will consist of a grid heat map overview with interactive axes. The heat map overview can be built using the CalendarView APIs[12] in D3. To create the interactive axes, we will also need to use a 1D CalendarView along with custom interactions in Javascript to allow the selection of the axes and category data matrix.

The last visualization will contain a line chart built from D3 Line Chart APIs. The challenge here is to dynamically create multiple line charts as the user is select states and display data from different dimensions (law, time of day, state, average, and etc).

[11] "Line Chart block - Blocks." 2013. 22 Oct. 2014 <http://bl.ocks.org/mbostock/3883245>
[12] "d3.js ~ Calendar View - Blocks." 2013. 22 Oct. 2014 <http://bl.ocks.org/mbostock/4063318>

8. Distribution of Work

This visualization can be split up into different modules and brought together in the end to implement interactions between each module. We have listed these modules and its sub tasks below and assigned them to team members. Note: these are proposed assignments of tasks. Depending on workload of each component, we will allocate resources as needed.

- Preprocessing of Data *[Siddharth]*
    - State Laws Data Compilation *[Siddharth]*
    - Data Normalization *[Siddharth]*
- Overview Line Chart [Saad]
    - Display average number of fatalities across US *[Saad]*
    - Implement selectors that span across the chart *[Saad]*
- Category Data Matrix *[Multiple]*
    - Implement Initial Category Map to build off of *[Saad]*
    - Display information between two categories *[John]*
    - Implement Interactive Axes [Siddharth]
    - Implement the selection of each bin in the data matrix *[John]*
    - Split boxes into two for each category in law mode *[John]*
- Detailed Line Chart *[Multiple]*
    - Implement initial line chart to build off of [John]
    - Display US Average [Siddharth]
    - Display state specific information when a state is selected [Hannah]
    - Dynamically add new line charts depending on the categories selected [Hannah]
    - Split off into two different lines for each state in law mode [Hannah]