# Visualizing Crime in Chicago

Zachary Fogelson, Alexander Jaffe

{zacharyfogelson, alexanderjaffe}@college.harvard.edu

## Overview and Motivation

The members of our project team are deeply passionate about building visualizations that use maps as an efficient visualization tool and also enabling individuals to interact with large data sets. We decided to focus on crime in Chicago because Chicago has a large crime dataset, >5,000,000 rows, which if not visualized effectively is unhelpful for people trying to achieve a general understanding of how safe different neighborhoods of the city are.

## Related Work

Many people are engaged in effectively presenting data which involves geospatial information. Consider, Mike Bostock's work on homicide in Chicago for the [New York Times](#). His depiction, of homicides in Chicago is an excellent representation of information which clearly allows people to see correlations between education, demographic information and murder rates. However, the visualizations is limited because it does not allow for user interaction. Similar work can be seen by Rob Paral through his blog, [Chicago Data Guy](#), and his company [website](#). Choropleth maps is a common tool used by Paral, but similarly he provides no interaction or dynamic filtering. Nate Silver's blog 538, combines choropleth with user interactions through the use of a tooltip
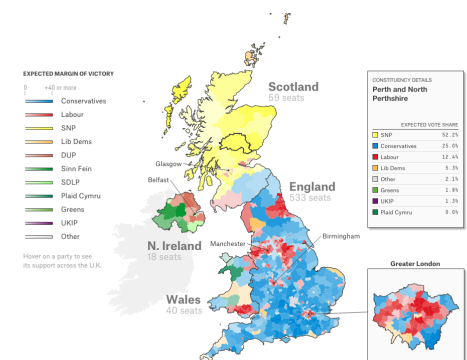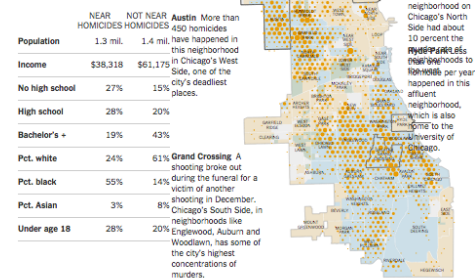
accompanying his map of the [UK general elections](#); however, he still does not facilitate data exploration because, his work does not include the ability to dynamically generate queries.

## Questions

Through interactions, we help users answer questions about the distribution of sets of crimes over the neighborhoods of the city of Chicago for a given time period. Furthermore, we expect to also be able to answer questions related to how different neighborhoods compare with respect to their crime rate, arrest rate, income, and demographics.

## Data

Our primary data source comes from the city of [Chicago's Socrata data portal](#).  Data are fairly easy to retreive using [Socrata Query Language (SoQL)](#); no scraping was necessary. Cleanup and wrangling, in our implementation, differ by visualization - we built a separate .js file (SocrataModel.js) to handle SoQL requests from our visualization components and hand them the appropriate data. Wrangling consisted mainly of reorganizing the SoQL response, which is in JSON format, to fit the particular needs of the visualization in question.
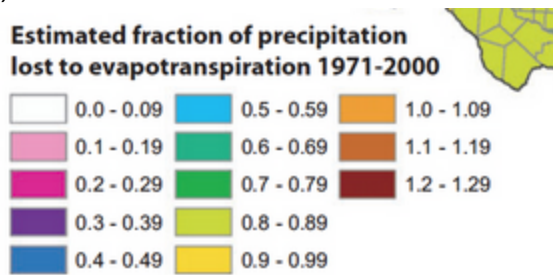
## Exploratory Data Analysis

Our exploratory data analysis consisted mostly of browsing the data Socrata's web platform, which allows the user to look at the data's categories and generate some basic visualizations. This more than anything gave us a sense for the size of the dataset (> 5M observations), revealing to us that a query paradigm would be more appropriate than having a local copy of the data. We also got a sense for the number of unique entries for our categorical data - important for assessing the feasibility of using the map and sunburst to represent them.
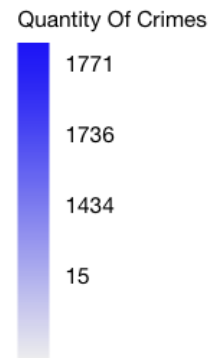
## Design Evolution

We knew we wanted to use a choropleth maps (see above). However, the representation of data on this map has undergone considerable evolution - we began with a quantized scale that sorted values into one of 9 ranges (similar to **a** below) and assigned them a color on a diverging color palette. Upon discussion, we realized the potential misleading nature of this scheme - community areas only 1-2 observations apart might appear drastically different if their values happened to fall over a quantum threshold. This led us to pursue the possibility of using a linear color scale. We feel this is a more accurate way to represent the difference between areas, and a better use of color as an encoding visual variable (**b** below).

a)

**Estimated fraction of precipitation lost to evapotranspiration 1971-2000**

| | | | | | |
|---|---|---|---|---|---|
| ☐ | 0.0 - 0.09 | ☐ | 0.5 - 0.59 | ☐ | 1.0 - 1.09 |
| ☐ | 0.1 - 0.19 | ☐ | 0.6 - 0.69 | ☐ | 1.1 - 1.19 |
| ☐ | 0.2 - 0.29 | ☐ | 0.7 - 0.79 | ☐ | 1.2 - 1.29 |
| ☐ | 0.3 - 0.39 | ☐ | 0.8 - 0.89 | | |
| ☐ | 0.4 - 0.49 | ☐ | 0.9 - 0.99 | | |

b)

Quantity Of Crimes

1771

1736

1434

15

We also underwent a design evolution process for our second visualization, the sunburst. The goal here was to represent part to whole relations of a nested 3-variable set of categories (crime primary type, description, and location). We considered a few options - mostly the treemap (**a** below) and a sunburst (**b)** - for visualization of these data.



a
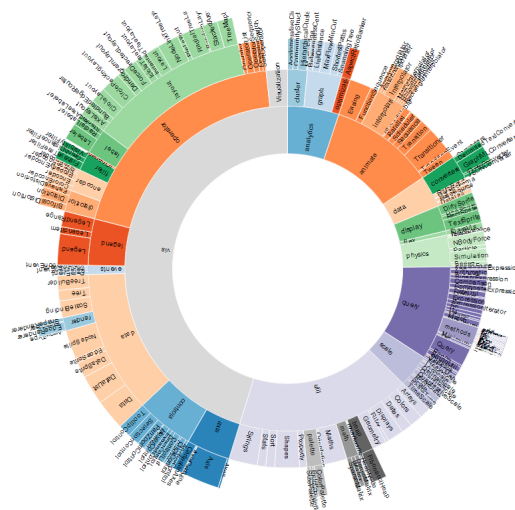


b

Both of these are easily implemented in d3. Ultimately, we decided to implement the sunburst - we feel it provides a more visually intuitive representation of part to whole and the nested nature of the data, a better user interaction. Interaction was crucial in our decision making here - while treemaps do have the capability for "zooming," the animation capability of the sunburst really allows users to feel as though they are "diving into" the data. Clicking on a particular segment leads to a new sunburst where that segment takes up 100% of the inner ring and previously deeper segments are now larger. To us, it is a natural representation of a hierarchy and will allow users to best understand that a third-level segment necessarily depends on the previous two segments in which it is contained.

Color in the sunburst was another thing we grappled with, particularly in the design studio - our peers helped us to realize that excessive categorical color in the sunburst was not just unnecessary but also potentially confusing. We ultimately opted for a sunburst where each tier has one color, reinforcing the hierarchical nature of the data and making identification of parents and children easier. Further, we implemented a mechanism so that the choropleth color corresponds to the color of the sun burst to make a visual link between the two visualizations.

We are also currently experimenting with a bar chart displaying arrest percentage and timeline which will be added into the overall visualization after Milestone 1.

## Implementation

The intent of our project is to provide a coordinated multi-view system for exploration of the Chicago crime dataset. The choropleth map will provide a geographic, contextual view of any data in question and will be linked with the sunburst, which will be used to navigate the dataset and dynamically update the map. For example, the user can select "ARSON" as a primary crime type option in the sunburst. This selection will trigger an event that updates the map, changing to display the distribution of arsons over the city during the selected time range. Ultimately, we'd like the interaction to be two way - selection of a particular community area on the map will trigger the sunburst to display statistics from that community alone. We will also eventually have a timeline that will provide a look at the selected statistic over the duration of our dataset, from 2001-2015. This will be brushable to dynamically update the data being displayed in the other two visualizations. All three elements are visible below.
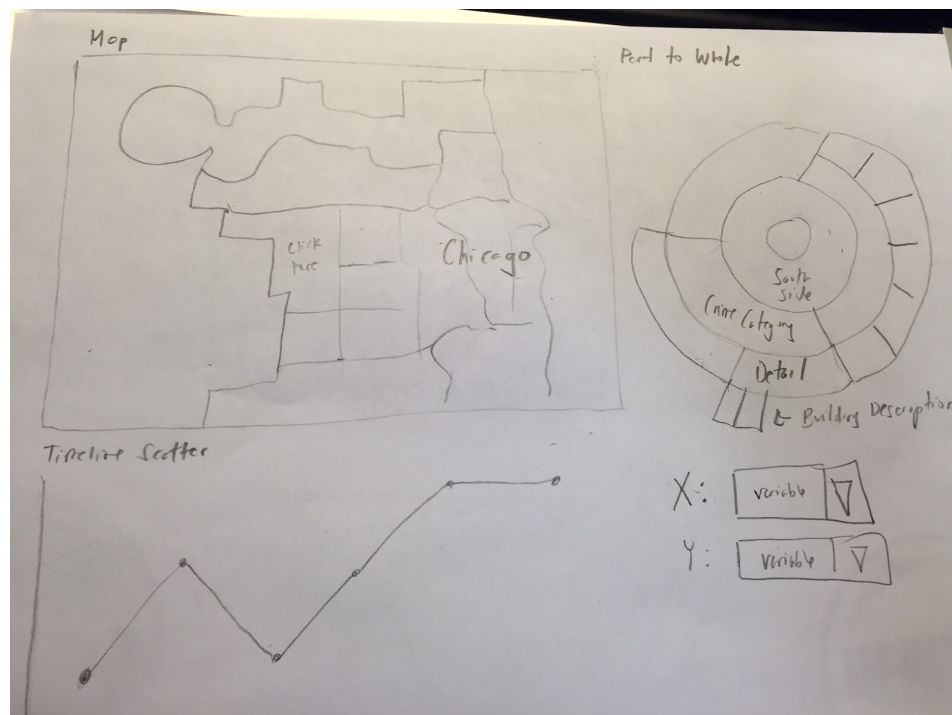
Finally, in order to allow for comparison between communities, we are interested in supporting pinning and stacking past queries to directly compare crime and demographic information between communities.

## Current Checkpoint

At this point in our project, we have implemented the sunburst for hierarchical data visualization and the choropleth for visualizing the quantity of crimes in a given area. We have also implemented the interaction between the sunburst and the choropleth shown below.
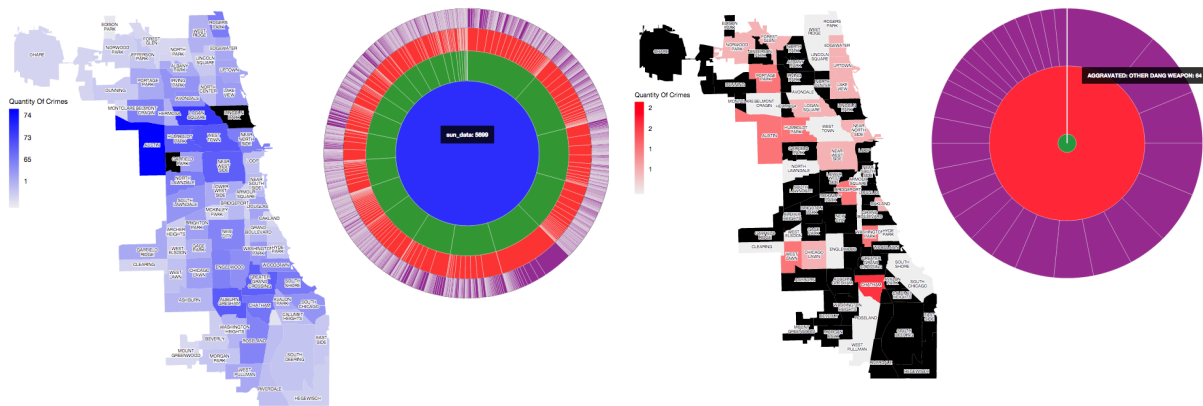


1) Distribution of all Crime
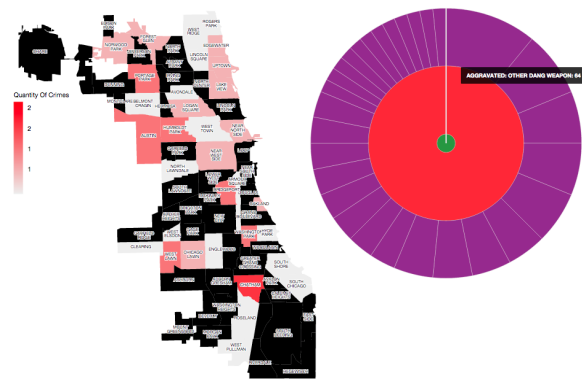


3) Distribution of battery with a dangerous weapon
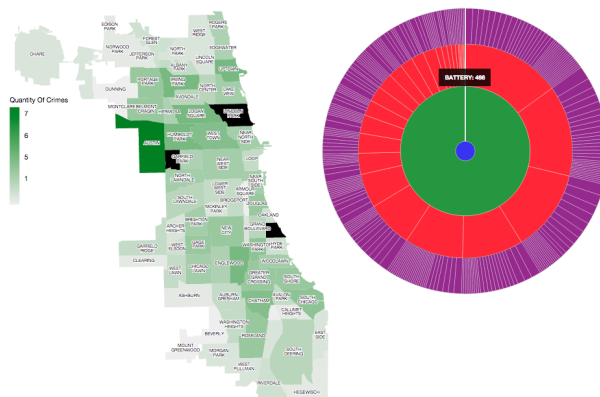


2) Distribution of battery



4) Distribution of battery with a dangerous weapon that took place in a residence

At each of the four steps above the user clicks on a ring of the sunburst which is one step further from the center. Each ring corresponds to a different category of information. Blue represents all crimes, green represents a category of crime, red represents the specific crime that took place, and purple represents where the crime took place. Therefore, in this example the user can see how many batteries took place with a dangerous weapon inside of a residence and where they were distributed throughout the city (it should be noted that the data used in this example is a subset of all crimes in 2004).

## Future Work



Going forward, we hope to implement a brushable line plot which will show the change in a given selection of the sunburst over time. Similar to a stock ticker, as seen on the left.

In addition, we hope to add a dashboard for each community which will show important information about that community in general. This is similar to the tooltip supported by the 538 visualization of the UK general election.

## Evaluation

So far, we've learned a lot from our visualization - where certain crimes are distributed in Chicago, where particular types of crimes happen, and how those trends are changing. We hope to highlight several interesting patterns as a narrative element in our final implementation.

The implementation is well on its way, but our future efforts will be directed to implementing the complex interactions between the components, optimizing performance with a large dataset, and adding the additional features described above.