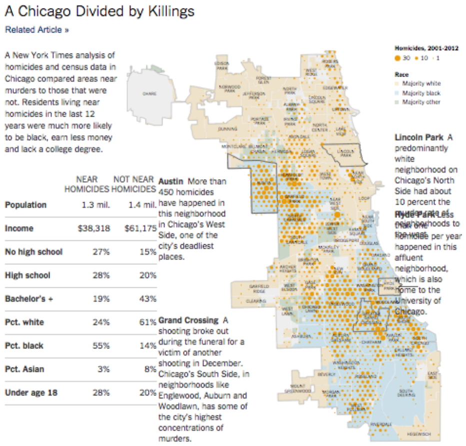
Visualizing Crime in Chicago

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.



A Chicago Divided by Killings

Related Article »

A New York Times analysis of homicides and census data in Chicago compared areas near murders to those that were not. Residents living near homicides in the last 12 years were much more likely to be black, earn less money and lack a college degree.

	NEAR HOMICIDES	NOT NEAR HOMICIDES
Population	1.3 mil.	1.4 mil.
Income	$38,318	$61,175
No high school	27%	15%
High school	28%	20%
Bachelor's +	19%	43%
Pct. white	24%	61%
Pct. black	55%	14%
Pct. Asian	3%	8%
Under age 18	28%	20%

Austin More than 450 homicides have happened in this neighborhood in Chicago's West Side, one of the city's deadliest places.

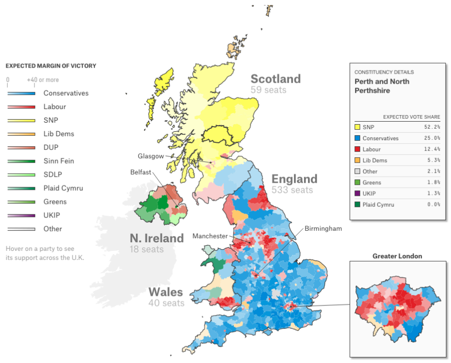
Grand Crossing A shooting broke out during the funeral for a victim of another shooting in December. Chicago's South Side, in neighborhoods like Englewood, Auburn and Woodlawn, has some of the city's highest concentrations of murders.

Lincoln Park A predominantly white neighborhood on Chicago's North Side had about 10 percent the murder rate of neighborhoods to the west.

Hyde Park Less than one homicide per year happened in this affluent neighborhood, which is also home to the University of Chicago.

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



CONSTITUENCY DETAILS

Perth and North Perthshire

	EXPECTED VOTE SHARE
🟨 SNP	52.2%
🟦 Conservatives	25.0%
🟥 Labour	12.4%
🟧 Lib Dems	5.3%
⬜ Other	2.1%
🟩 Greens	1.8%
🟪 UKIP	1.3%
🟩 Plaid Cymru	0.0%

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.

N.B. As we dynamically queried the Socrata database in our project, the crime data that we used for the bulk of our visualization is not included in the repository. However, demographic, income, and mapping data are included.

We decided to dynamically query for several reasons: 1, to reduce computational burden, and 2, to allow our visualization to effortlessly and automatically update with time as the data update.

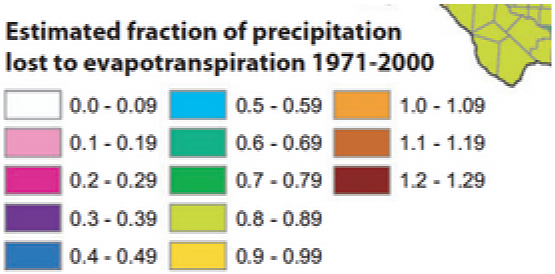
Exploratory Data Analysis

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.

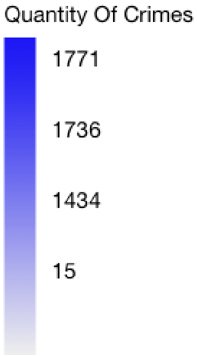
Design Evolution

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).

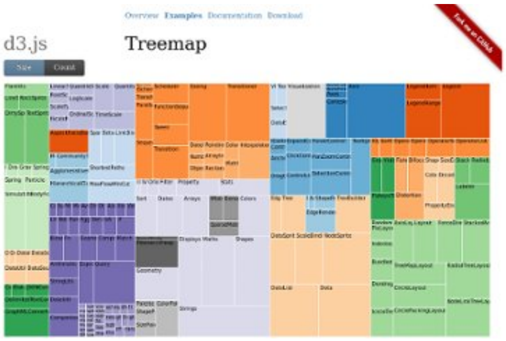
a)



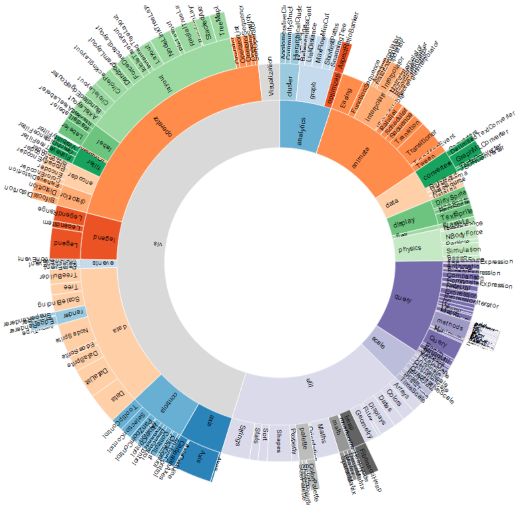
b)



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 also considered a sunburst where each tier has one color, reinforcing the hierarchical nature of the data and making identification of parents and children easier. Ultimately, however, we implemented a hierarchical scheme that assigned a set of hues to each primary subdivision, visually reinforcing fundamental splits and aiding in user interaction. 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.

Our timeline underwent significant design evolution throughout the process. Our goal was to show changes in a particular crime type on a temporal scale. Our first question was about the degree of data aggregation - how much resolution do we want in our timeline? A years-scale promised higher counts but a lessened ability to pick out sub-trends. Ultimately, we decided to leave the timeline at the finest possible resolution - days - as it requires the least amount of modification from the dataset and is most intuitive for brushing.

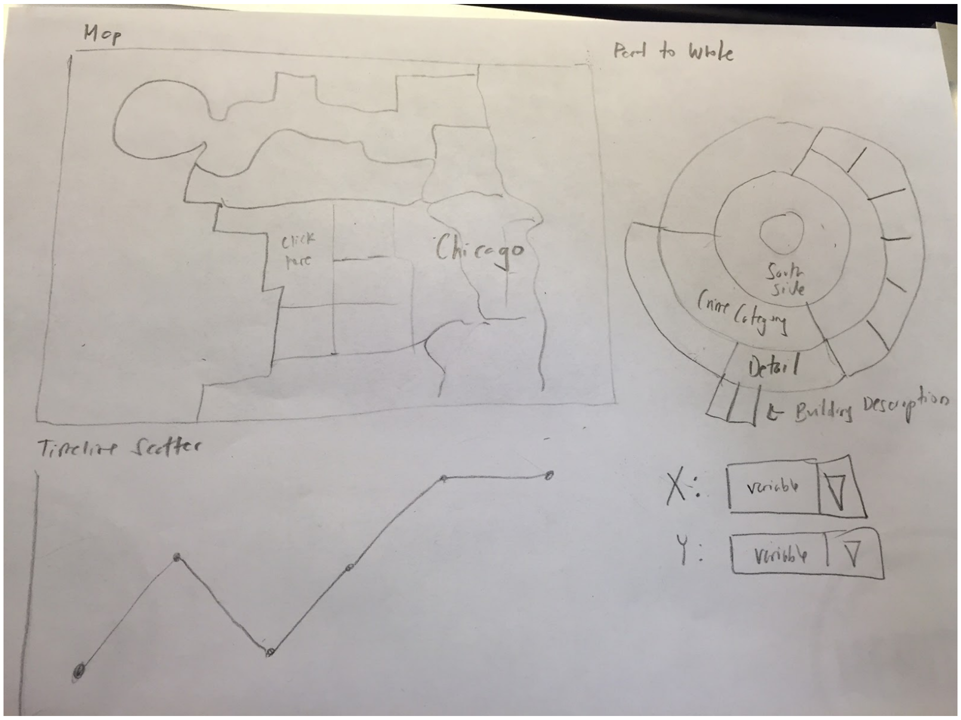
Another visual challenge was the type of shapes to use on the timeline - we experimented heavily with the path svg (as in Homework 3). However, this had a few issues - transitions were unwieldy and confusing - the path would cross over itself - and, at lower resolutions, felt misinforming - paths naturally show a degree of interpolation between the actual data points. For this reason, we decided to use points and a trend line so that users can see where the actual data points are while also visualizing trends. We believe this is a more honest and accurate representation of the data.

Finally, we also had to deal with missing data in the timeline - crimes are only logged when they occur, not when they don't occur. To deal with this, we created a zeroed data frame of the appropriate size and then dynamically filled it in using the data. This way, time slots with no crimes logged are actually shown as 0 instead of NA.

Implementation

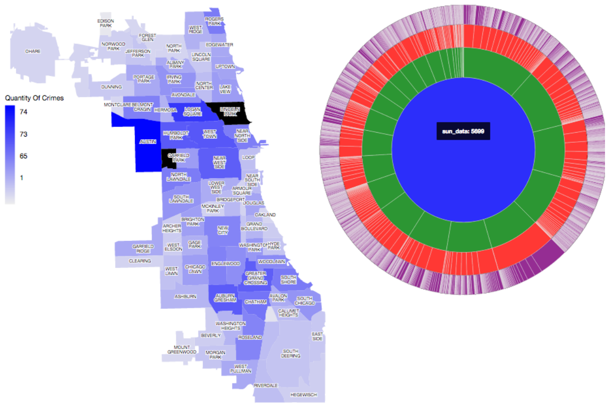
The intent of our project is to provide a coordinated multi-view system for exploration of the Chicago crime dataset. The choropleth map will provide a geographic, contextual view of any data in question and will be linked with the sunburst, which will be used to navigate the dataset and dynamically 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. Our timeline provides a look at the selected statistic over the duration of our dataset, from

2001-2015. This timeline is brushable to allow dynamic updates the data being displayed in the other two visualizations. All three elements are visible below in our primary rendering.

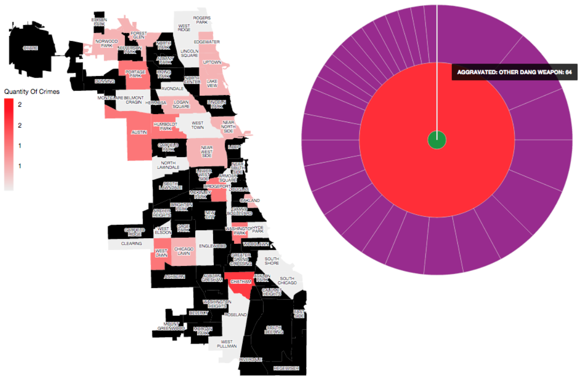


Milestone (April 17, 2015)

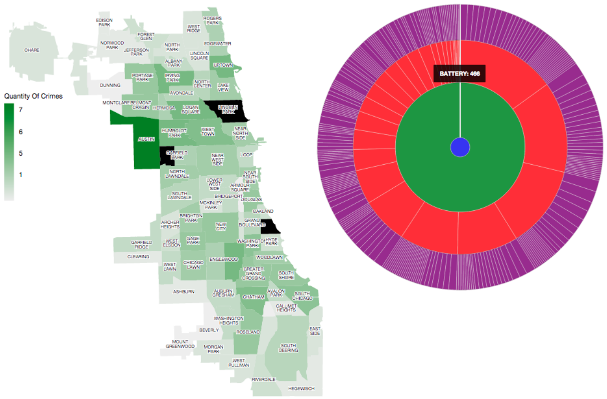
At this point in our project, we have implemented the sunburst for hierarchical data visualization and the choropleth for visualizing the quantity of crimes in a given area. We have also implemented the interaction between the sunburst and the choropleth shown below.



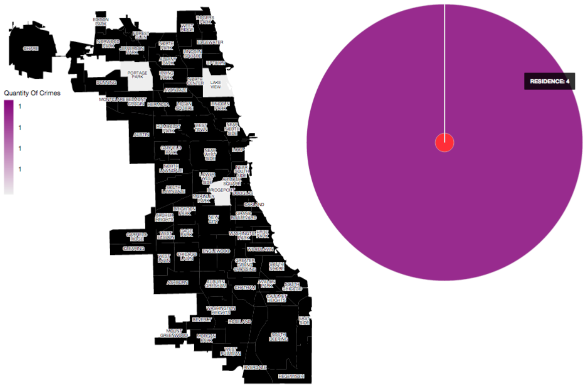
1) Distribution of all Crime



3) Distribution of battery with a dangerous weapon



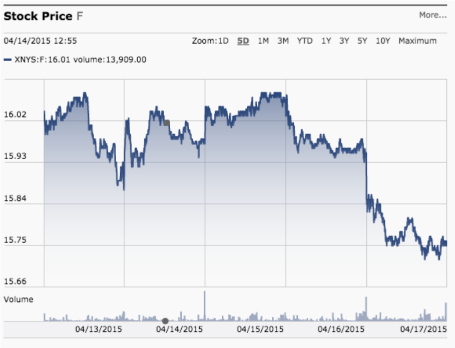
2) Distribution of battery



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.

UPDATE (5/5/15 FINAL Submission): We did ultimately accomplish the brushable line plot and community dashboard mentioned above in our milestone report.

Evaluation

Ultimately, we 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 were able to highlight several interesting patterns as a narrative element in our final implementation.

Through our process we also learned about the nature of user interaction and design iteration.