# Dallas Recorded Incidences Report

Guojun Wu, Qian Lu, Xinyu Dai
{gwu, qlu, xdai4}@wpi.edu
{github:beflew, qianlu-wpi, SarahDai}
Data Visualization Final Project
CS573(Fall 2015)

**1.Overview and Motivation:**

On the evening of Nov.13th, 2015, a series of the coordinated terrorist attacks occurred in Paris. Under today's global environment, the need for good record-keeping and information-sharing practices in incidents has taken on added significance. Thanks to this big data era, we could take full advantage of these huge incident data sets. Not only do good records provide crucial internal information, law enforcement agencies could communicate agency-to-agency and across continents based on these records and analysis in order to protect the Nation's citizens. Nothing is more important to accomplishing that mission than having accessibility to accurate and timely records. Calls for service records and investigative, arrest, criminal identification, detention, and even civil records hold information that by themselves mean little; however, when pieced together with information from other jurisdictions, the result can help with all levels of investigations and aid in safeguarding the Nation.

Our team focus on the RMS--Records Management System of incidents reporting in America. A RMS consisting of uniform records in incident/offense reports is extremely beneficial for law enforcement agencies, which satisfies operational and management needs at the local level, as well as the informational requirements of FBI. We searched the open RMS data of several major cities like Boston, San Francisco, etc. and chose Dallas police reported incidents dataset with 135,285 instances.

We chose this data because of the detailed attributes for each incident like the description of the offense type and most of the the time when the incident happened is quite precise. Columns concerning whether the offense involved drug or gang could offer more interesting information for our data analysis and visualization.

In our initial proposal, we proposed that the intended audience for this visualization will be primarily people who want to know which area in Dallas has less incidents of a specific type, which potentially indicates the safeness of that neighborhood, though perhaps additionally for the more specific subsets of governments, insurance companies, etc. After collecting the feedbacks, we decided to narrow down the intended audience/users to only people who want to know which neighborhood is safe to live.

**2. Related Work:**

We visited several existing crime mapping websites, most of which contain general descriptions about the crime datasets. Consider, one of the projects from Harvard University Visualizing Crime in Chicago. Their district-based analysis gave an excellent visualization on the correlations between the household income, demographic information and quantity of the crimes. Another example of the Dallas crime map from http://spotcrime.com/tx/dallas marked all the locations where the crimes ever happened in the map. This might be too specific which would be hard for the users to find several patterns or generate useful suggestions.

**3. Questions:**

In our initial project proposal, we would like to answer the following questions:

- Can we find a geographic pattern on the map such that which neighborhood/area in Dallas has relatively more or less reported incidents than the others during 2014/2015?
- On a given date, how many incidents are reported in total? Among these incidents, how many incidents are car accidents, robbery, suicide, etc?
- On a given date, which time frame is likely to have more reported incidents? Is this a pattern on a daily basis?
- Given a period of time, how many incidents of a specific type were reported? Where did they happen?

Based on the proposal feedback and the presentations from the other groups with similar crime datasets, we decided to focus on answering the “where” questions: to give a detailed description of safety degree for each part of each street. We tried to give the general public a useful reference information similar like on-time traffic flow information. Moreover, we want to divide the offense started time into several time intervals like day and night and midnight, weekdays and weekends, etc. to figure out if there’s any pattern in the time frame. This might be helpful for the government operation.

**4. Data:**

**4.1. Data Source**

We acquired and used data of reported incidents from Dallas Police Public Data website. The data set contains reported incidents from Jan.1, 2014 to Nov.17, 2015. The link is https://www.dallasopendata.com/Police/Dallas-Police-Public-Data-RMS-Incidents/tbnj-w5hb

We could use the open-source library for interactive map from http://leafletjs.com/

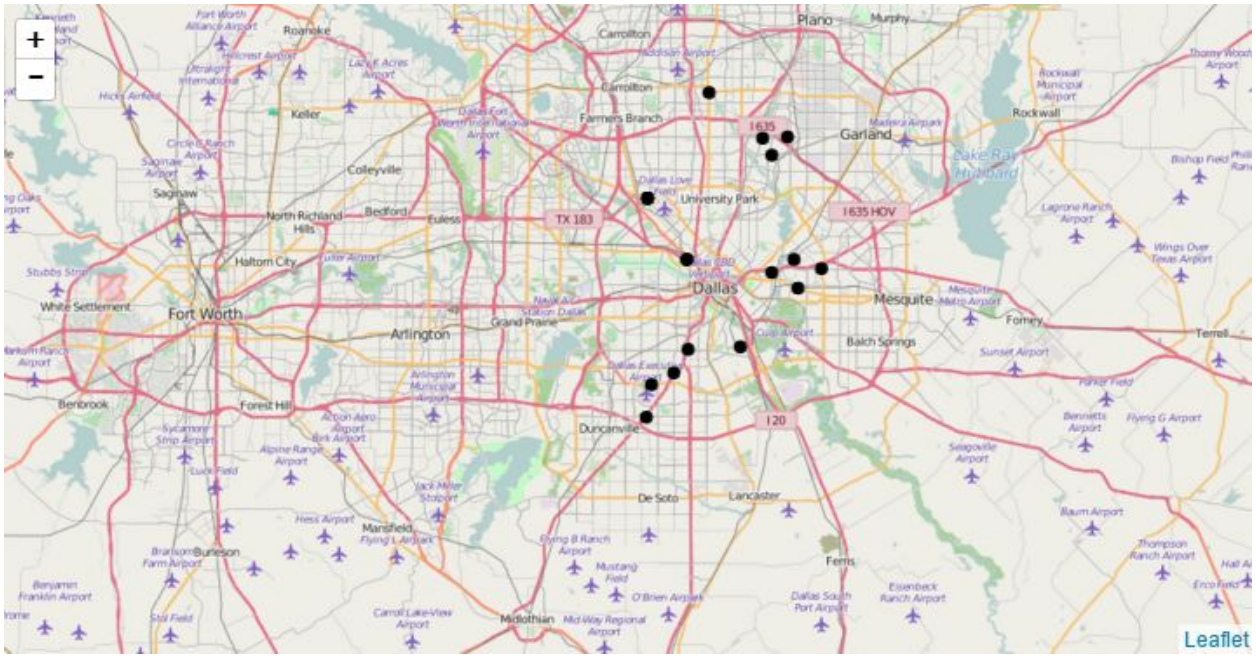
### 4.2. Data Scraping and Cleanup

After collecting the dataset from Dallas Open Data, we found that the dataset is pretty large (122MB), and there are lots of attributes which are irrelevant to our project. Therefore, we extracted 15 of 100+ features. By doing that, we reduce the size of our dataset down to 25MB.

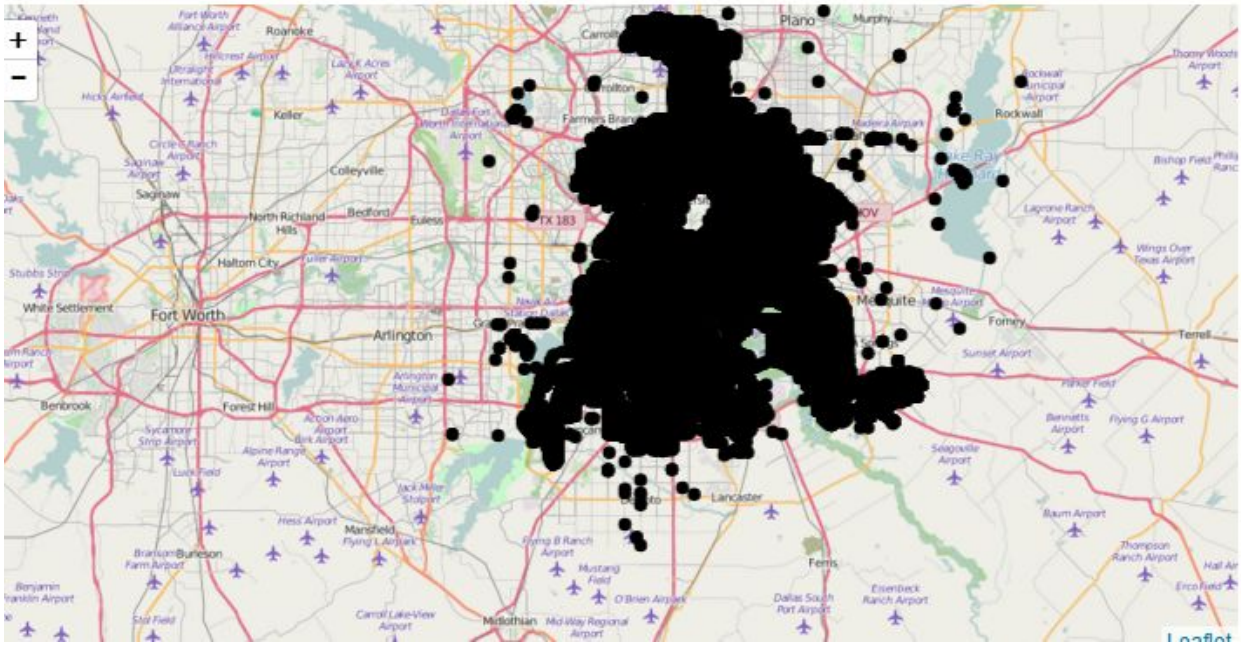
But it's still not small enough, so, we decide to split our dataset into multiple datasets. For each visualization, we use only attributes that we need to implement the visualization. Then, we can have a metadata file to connect all the visualization. Moreover, we remove all tuples that contain null values.

## 5. Design Evolution:

At beginning, we have projectioned all the incidents to a map, which is shown below.

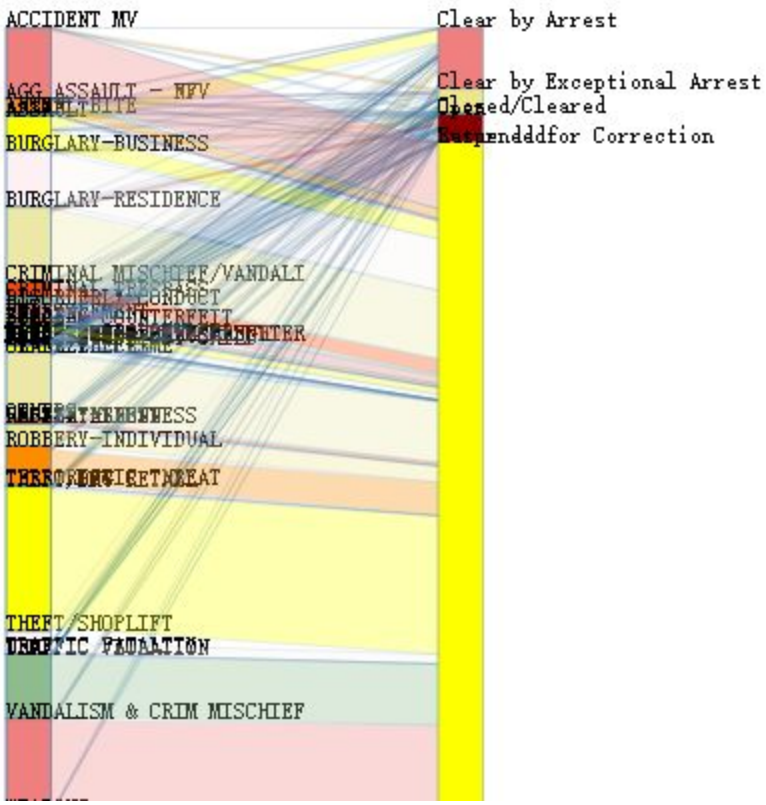



But when we loaded all data we had, we found it not only slow but also pretty ugly and unclear.

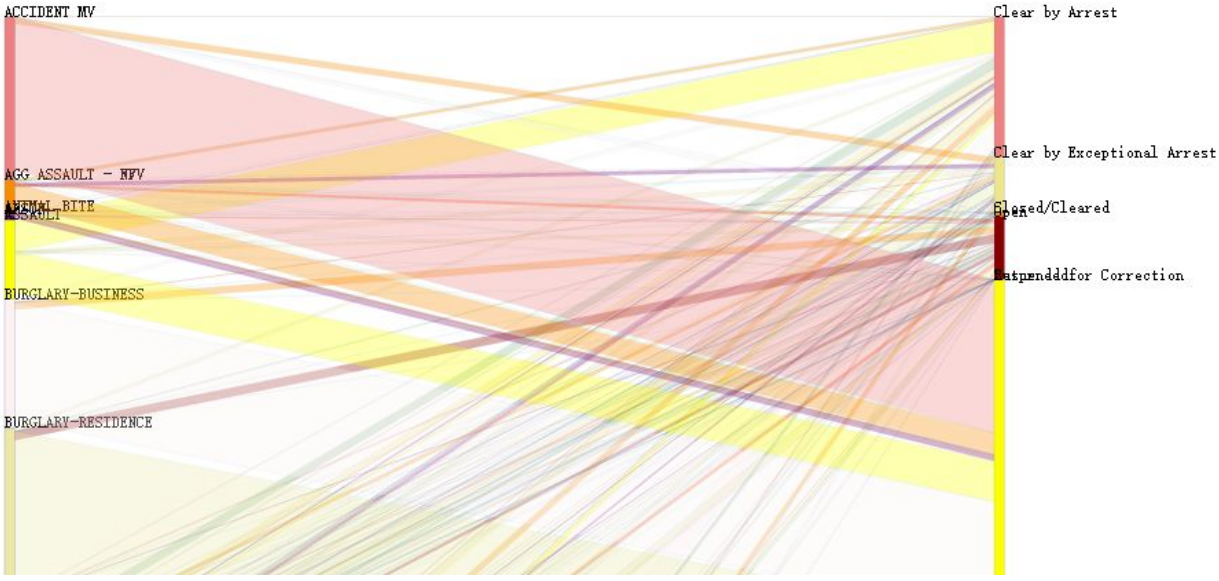



So, what we decide to do next is to rebuild this map using a different way. For now, what we have in mind is, using opacity, using sampling, using different data for different zoom level or, the most exciting one to us, using street-level heat map, which is pretty like a traffic map to show our data.

Also, we use a BiPartite to show results of different type of offense incidents.

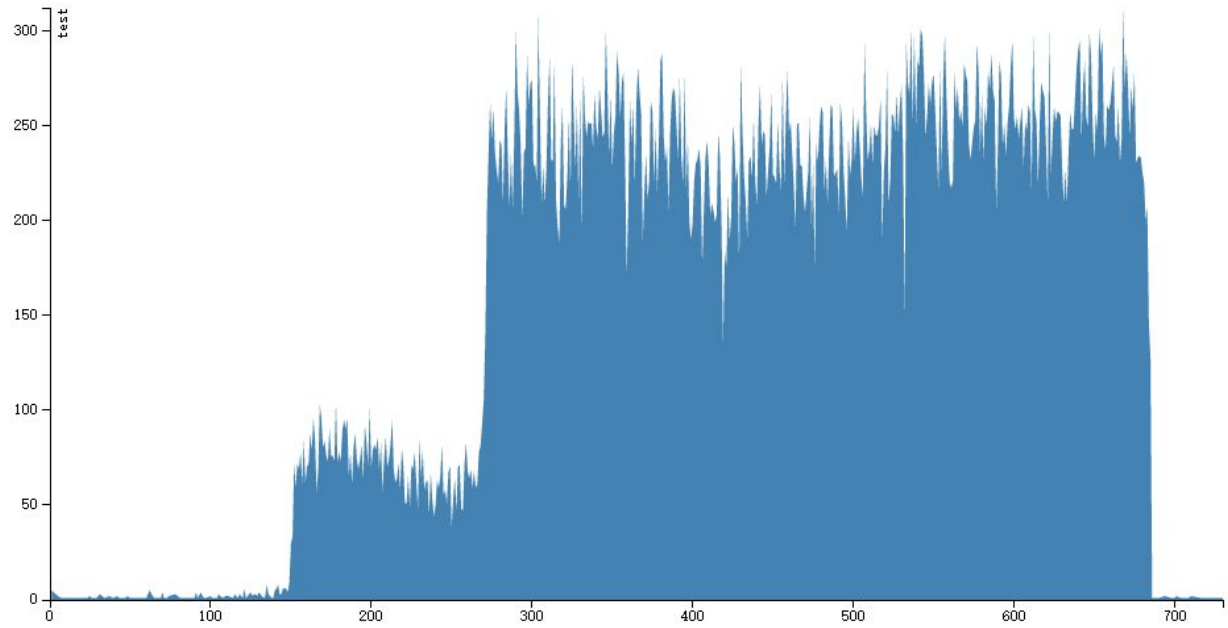



Also, we have a bigger picture like below…:



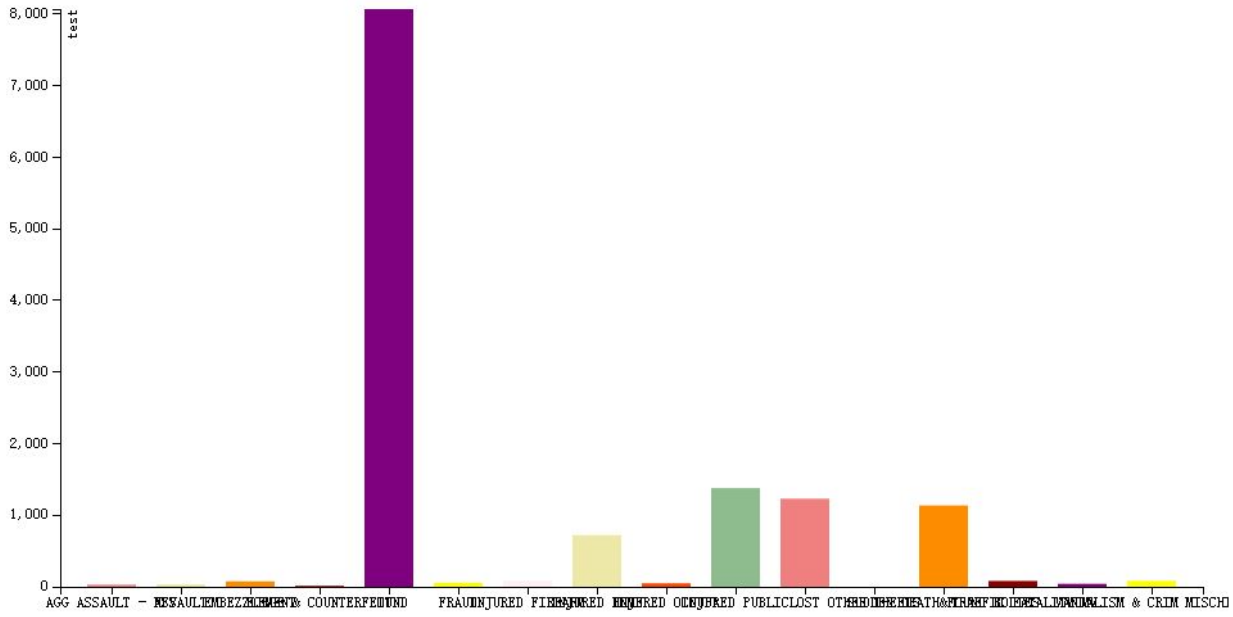

The problem we have here is that we need to make this chart more clear to audiences to know what we are going to convey.

The third visualization prototype we have implemented is an area chart.




From this chart, we can find out that we don't have enough data for the first 280 days of 2014. So, when we talk about "when" question in the next few weeks, we should try to avoid using data from the first 280 days of 2014.

Then, we have a bar chart for amount of different types of non-offense incidents.




In this chart, some categories can hardly be seen. That's what we should fixed.

**Work Log**

**Week 1**

During the first week, we modified our project slightly based on the feedback of our proposal. Then, we assigned work for each of us. By the end of first week, Xinyu collect all the data we need and Qian did the preprocessing. As for Guojun, he came up with the idea that we can split datafile into different datasets.

**Week 2**

In this week, we started to implement our thoughts.
Guojun did the map part, Xinyu and Qian did the rest three charts.