Spatial-Temporal Crime Data Analysis and Visualization for Los Angeles, California

ShuHo Chou, Kairi Kozuma, Dennis Sosa, Chiamin Wu, ChiLin Wu

Project Report
CSE 6242
April 24th, 2018

I. Introduction

In recent years, Los Angeles (LA), California saw an uptick in crime. According to [4], violent crime rose for the fourth year in a row. LA residents, police, tourists, and policy makers may want to know time and locations that are safe. This project will implement a data analysis and visualization tool to solve this problem.
A summary of reasons why this may be useful to particular people is given in Table 1.

Person	**Reason**
Tourists	Avoid dangerous areas
Police	Plan patrol routes
Residents	Change commutes
Policy Makers	Allocate resources appropriately

Table 1. Reasons to use the project

II. Objective

The objective is to visualize and predict LA crime data on a website, with the following features:

- Point plot of crimes
- Filter by crime type and victim traits
- Filter by time interval
- Suggest dangerous areas to avoid based on predictions

III. Literature Survey of Current State and Limits

Current crime maps lack temporal filtering and makes broad assumptions, combining all types of crimes as a single measurement. This project aims to remove those assumptions by letting users use advanced filtering techniques.

A. Crime Mapping

The US DoJ summarized modern crime maps in [5]. Point mapping (A in Figure 1) was the most common technique; however, there is difficulty interpreting large datasets [5]. This project will create a map with sophisticated methods, such as gradients.

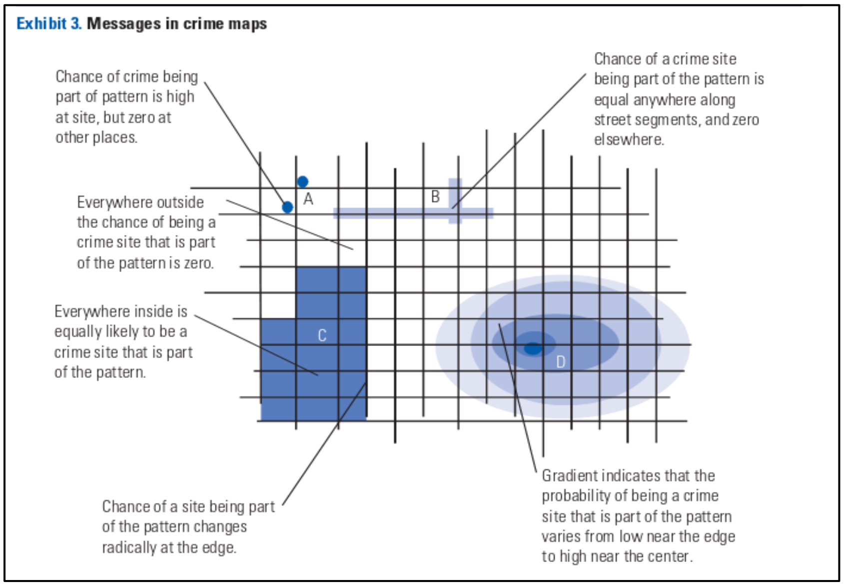


Figure 1. Various crime-mapping technique

Researchers in [16] plotted crime distribution per hour (Figure 2), using temporal residual networks; however, usability is poor without an actual map. This is useful since this project will use temporal filtering, and improvements can be made by overlaying an actual map.

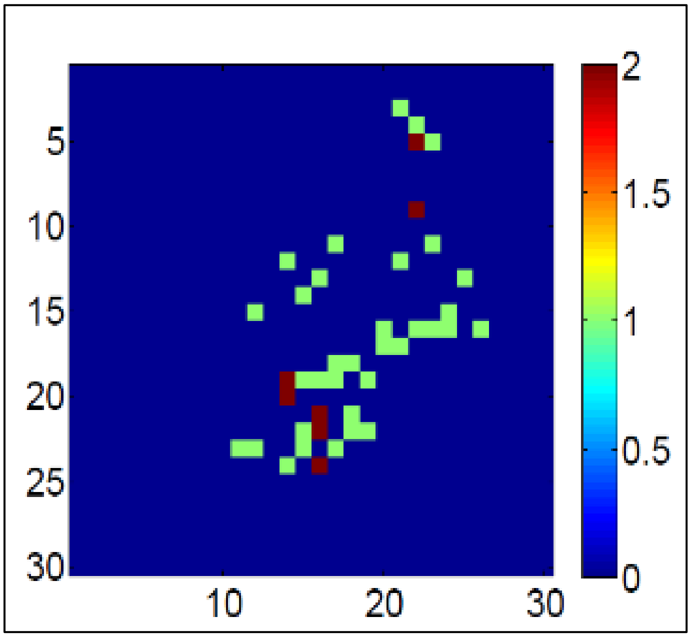


Figure 2. Crime distribution at 1:00 AM for LA

[3] describes novel crime mapping methods, such as temporal directionality to correlate crime time and locations; however, it is difficult to interpret without expertise. This project will use time filtering to uncover temporal directionality and present them in a friendly interface.

[7] combines spatial and temporal information together to explore crime patterns within socio-economic and environmental neighborhoods. This project can use such methodology to analyze crime behavior and determine causes, however, this paper only considers burglary, and we can include more types.

[14] proposed a geographic, hierarchical self-organizing map to analyze georeferenced tweets with geospatial and temporal meanings. This project can use this method to facilitate determining dangerous times and places to avoid. Improvement can be done by including socio-economic information as stated in [7].

B. Crime Predicting

Current crime prediction models use supervised learning. In [1], decision trees and naive Bayes classifiers are used to predict crime space and time, but accuracy rate is low. In [2], random forest classifiers are used to make high-accuracy predictions of the influence of urban indicators on homicides, but it tends to overfit. For these two papers, we can experiment with urban indicators and we can employ neural networks to make our crime predictions while being mindful of any possible overfitting.

[12] uses statistics to extract likely features from criminals, which might be useful for phenotyping. But the shortcoming of [12] is that some features might not be useful and should be avoided (ex., scammers are in scamming groups).

[6] uses phone records to distinguish and visualize the criminal networks. We can leverage this heuristic to build our classifier. The shortcoming of [6] is personal records are hard to get, and might not be helpful for people to avoid dangers.

[11] uses arrows to visualize relationships between victims and offenders, which can help people understand their relationship. Our project can also deploy some simple geometries like this. The potential drawback is that colored-dots are hard to differentiate between victims and offenders.

C. System Implementation

[13] describes a framework for crime mapping, which follows an MVC paradigm [13]. The project will mimic this approach, since the project will be hosted online. This project will improve upon the paper by incorporating temporal aspects to visualization.

The Atlanta Data Dashboard (ADD) introduced in [8] uses visualization tools for public safety.

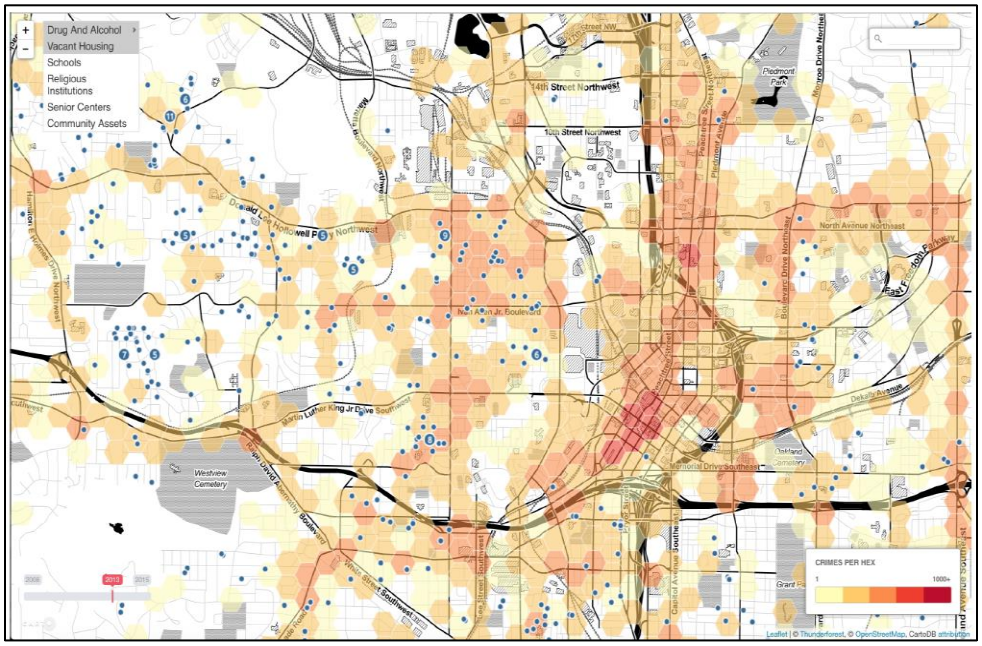


Figure 3. Spatial Visualization of Crime Data as Hex-based Heat Map within ADD [8]

[8] can help us implement our heat map plots. The ADD does not contain much user interaction or filtering, it just visualizes crime data.

CrimeProfiler uses many natural language processing techniques to extract and curate crime information from news media [15]. We can use these techniques to process crime descriptions and categorize crime types. CrimeProfiler visualizes crime statistics using bar graphs and pie charts. Our system will visualize both statistics and geographical regions with prediction techniques.

D. Clustering Data

CrimeVis extracts criminal data to visualize it in multidimensional, spatiotemporal, and multivariate graphs [9]. It uses statistical analysis and clustering techniques to gauge criminal trends. The team can use these techniques to visualize the LA crime data. While CrimeVis visualizes geographic states as a whole, we will visualize the city of LA with specific coordinates.

[10] visualizes crime data by showing offender and victim ages against number of incidents; however, this was not obtained automatically. We can improve this by obtaining information through supervised learning, with filters for gender, district, and time.

IV. Proposed Method + Experiments

The project approach is novel, because it focuses temporal and crime type filtering on top of traditional crime mapping techniques. Online crime maps do not have such capability; our system's approach can benefit LA residents and tourists. Users can distinguish different crime types and the locations where these crimes occurred. Additionally, our system contains a prediction model that can predict the areas where a crime is most likely to happen depending on the day of the week and the time.

A. User Interface - Map

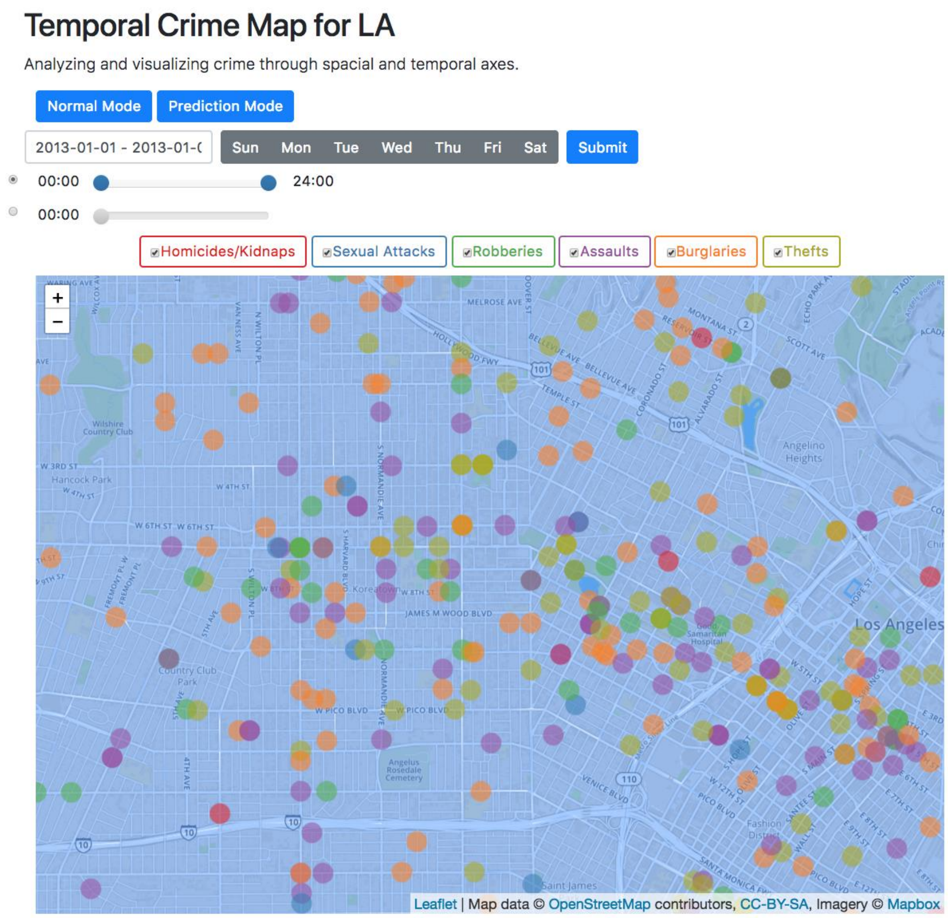


Figure 4. User interface of the crime plotting website

Figure 4 is a screenshot of the current user interface of the web application. The map can be scrolled and zoomed in or out. Each incident from our dataset is displayed on the map as a colored semitransparent circle, with opacity at 50%. The circle sizes are scaled proportionally with the map.

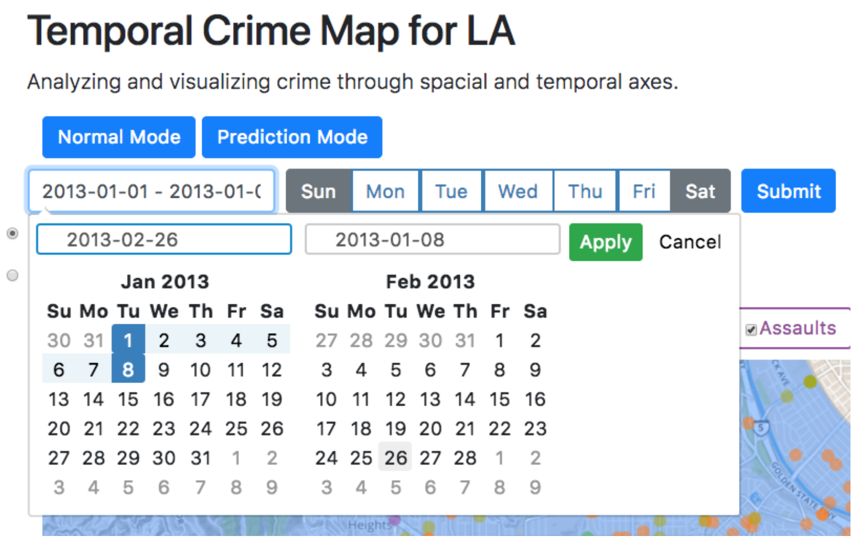


Figure 5. Date filtering capabilities of the website

All date filtering capabilities are present on this interface. Figure 5 shows the date and day of week filters. Day of the week can be chosen by selecting or deselecting each cell. In Figure 5, Sunday and Saturday were chosen to show incidents for the weekend. These values are used to filter incidents of a database, triggered by the "Submit" button.

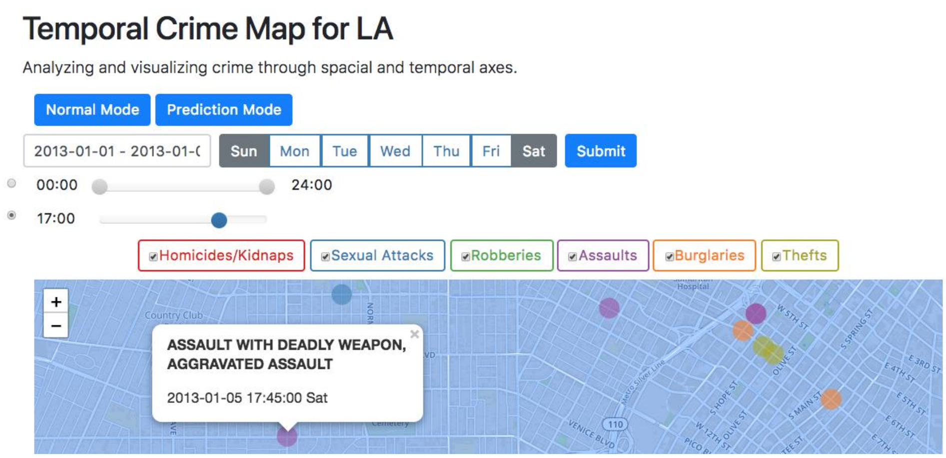


Figure 6. Time filtering capabilities and popup of incident

In addition, there are two types of time filtering. One is a range filter, specifying both the start and end hour for crimes. This is shown as the top slider in Figure 6. The other is an hour interval filter, which shows all crimes within the hour, shown as the bottom slider in Figure 6. This can be used to easily show the trend of crimes over time, by sliding across the filter widget. Clicking on an incident circle brings up a popup that shows the crime description, date, and time of occurrence.

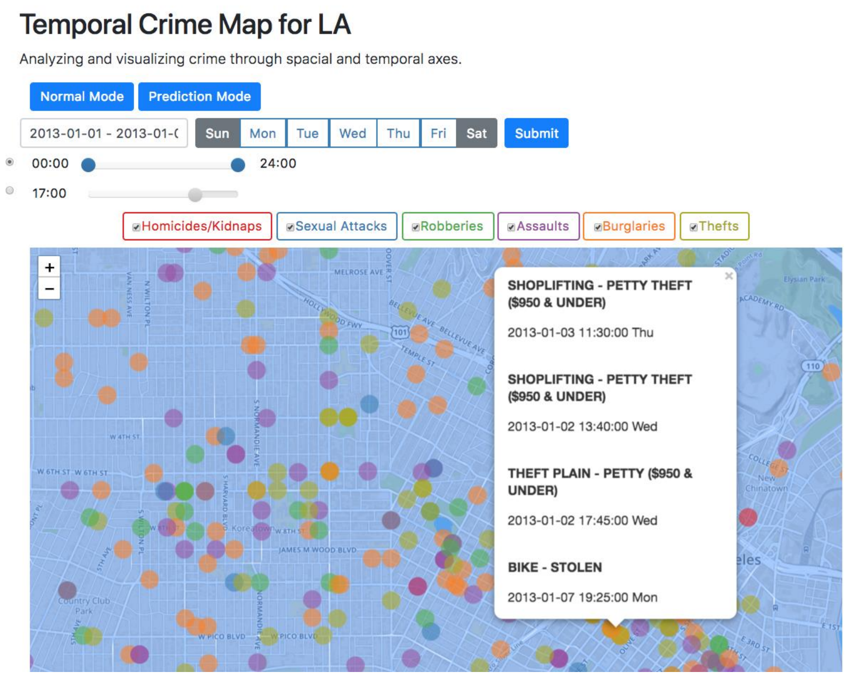


Figure 7. Multiple crimes at same data point

There are crimes that occur at the same data point. To allow users to view these crime types smoothly our approach makes sure to keep track of each data point's longitude and latitude in a dictionary in JavaScript to plot each data point with the appropriate crimes as shown in Figure 7. The checkboxes are part of the filtering as shown above. We use the "Submit" button to control a checkbox, letting users define their search range easily. To speed up filtering computation, we filter out unnecessary data beforehand. To increase interpretability, we use six colors to represent six different crime types without losing data completeness. Users can visualize coarse category by color and fine description after clicking on a specific point. The colors are assigned based on seriousness. The darker the color, the more serious the crime.

B. User Interface – Google Data Studio Dashboard

The bottom half of our web app is the Data Studio (DS) dashboard, a beta product of Google Cloud Platform (GCP). First, our original csv data was imported into Data Storage. Then we used BigQuery (BQ) to setup a dataset, to import the file from Data Storage, and to clean the data. After that, we moved on to DS, imported the dataset we cleaned from BQ, and tried out different possible user interfaces in DS.

During our experiments on DS, we have discovered that although DS is Google's analytics product and thus have great compatibility with other Google products, DS does not provide the aforementioned street level map function. Contrarily, DS only shows geo-maps on a state level. It further confirms that our app compliments or even improves current GCP product.

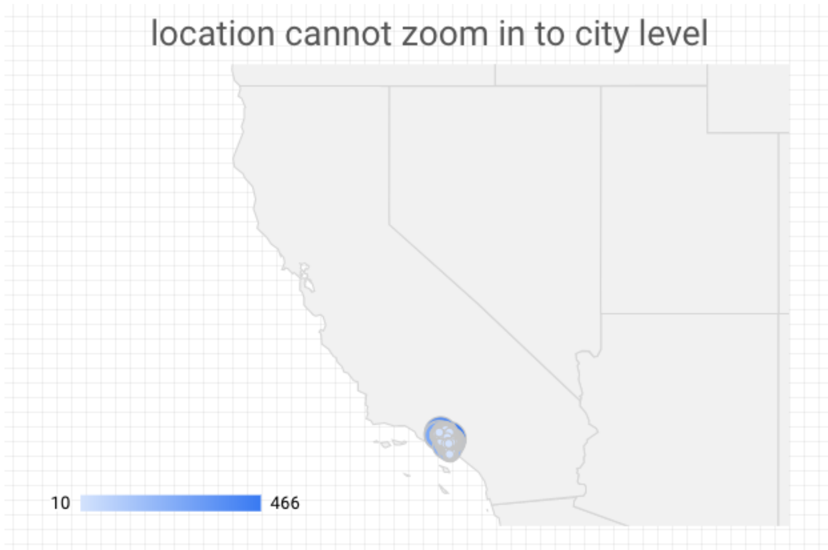


Figure 8-1. Data Studio map function only on state level

Our Data Studio dashboard intends to show statistics of crime counts by hour, sex, age, day of week, ethnicity, crime type, and date. It complements our map by showing users the statistics and the composite among different attributes. Users can easily filter by dates, hours, and types of crime on top of the dashboard, and the dashboard will quickly query and show the crime counts verses various variables interactively to the user.

After some testing of the dashboard layout, we found our final layout below fits the purpose and complements our map the best. At first, we tried the layout by putting all these 8 components on different pages. While the page looks cleaner and less crowded, users might have to switch between pages to read from different components. This greatly reduces the readability and the comparability of our graph components. Furthermore, due to the fact that DS filters are on a page basis, if users switch to different pages they will need to manually select filters again, which makes the dashboard cumbersome to use. Secondly, as the aforementioned map that we built is a temporal and special spectrum, we want our users to understand the crimes in LA by showing them the breakdown of the graphs. We experimented a few graph options and found the following to be the easiest for users to understand:

1. Bar chart to show crime counts by hours, crime counts by sex and hours, average victim age by hours, crime counts by day of week, and crime counts by types of crimes
2. Pie chart to exhibit composition of victim sex and victim ethnicity
3. Time series chats to display crime counts by date

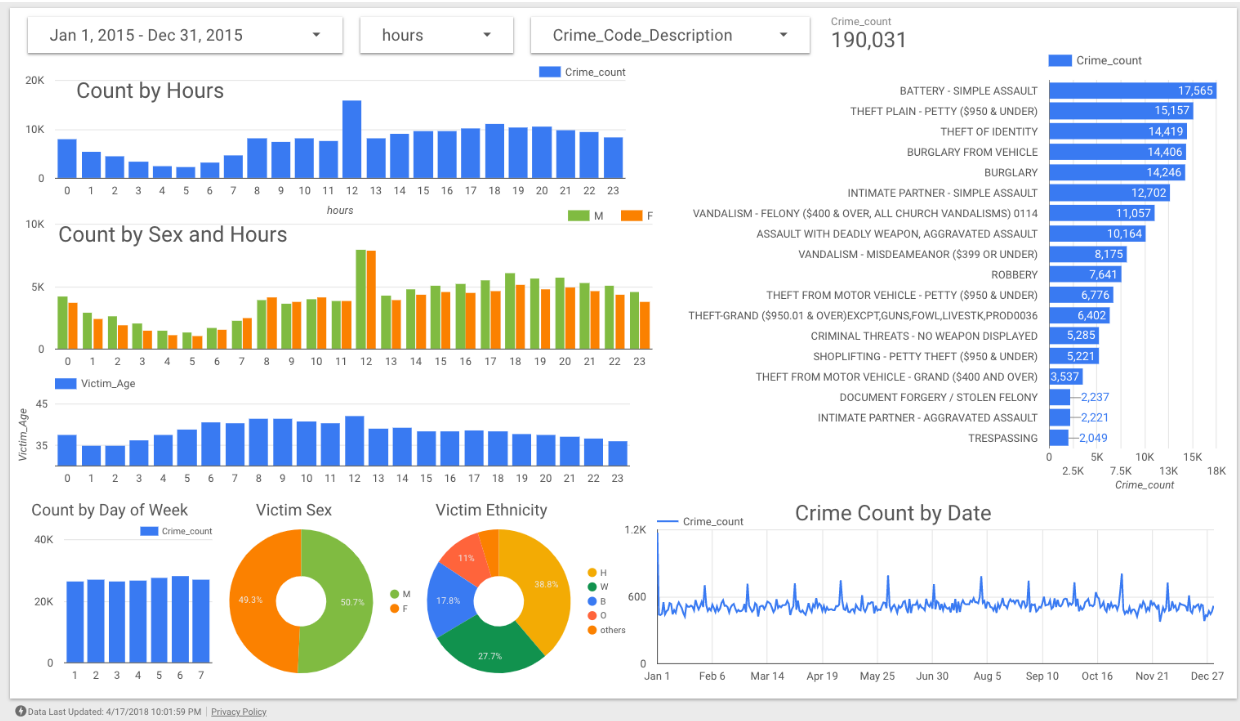


Figure 8-2. Data Studio interactive user interface

C. Observations

While retrieving and plotting 1,000 points is fast on modern browsers, raising the number of incidents to over 5,000 results in significant slowdowns. One solution is to randomly sample from the set of filtered endpoints, such that the number remains below a certain threshold. This keeps the relative frequencies of crimes based on location to be the same; however, this does not shorten the retrieval time and it distorts the actual crime plot data. More sophisticated methods of combining incidents before plotting were experimented but we chose to randomly sample.

D. Criminal Event Prediction

After visualization, we further improved our results. In this case, we built a dynamic criminal event prediction function. In this function, we gave weights to each criminal type at first. Then, we calculated the dangerousness scores and put it into our prediction model. In this model, we browsed historical data as we showed on our map, took the average of the dangerous scores and provided a final score to show our users how dangerous a particular region is during their selected time zone. In this case, it can help users easily understand our map and increase the feasibility.

We separated LA into 1135 shapes based on the LAPD district shapefiles. For each district, the dangerousness score is defined as the sum of occurrences for each type of crime multiplied by their corresponding weights (shown below) in a certain amount of time (say within one month). To train a model that can predict how dangerous all the districts are, for each district, we get the level of danger for each day in the past *window_length* days as one training data instance, and the sum for the next *label_window_length* days as the corresponding label.

CrimeCode	Description	Weight
1	Homicides/Kidnaps	100
2	Sexual Attacks	60
3	Robberies	40
4	Assaults	30
5	Burglaries	10
6	Thefts	5

Table 2. Weights for each crime type

Below is a small experiment of our crime prediction model.

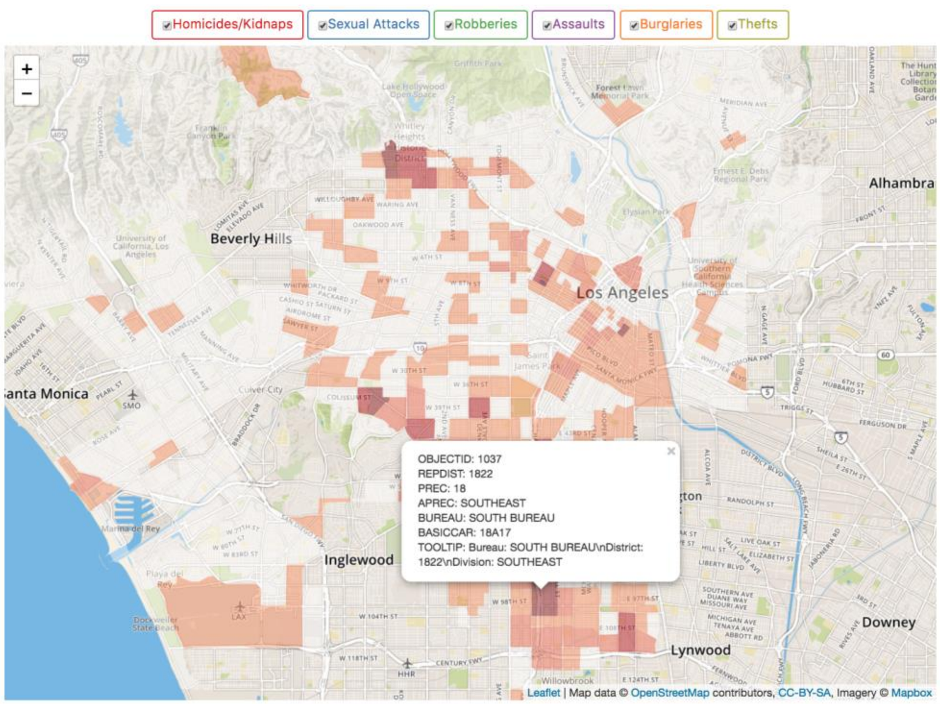


Figure 9. Criminal event prediction experiment

As shown above, our trained prediction model predicts areas where crimes are most likely to occur based on the training data. As previously mentioned, the darkness of a zone's color indicates the level of danger.

V. More Experiments

A. Visualization Experiments:

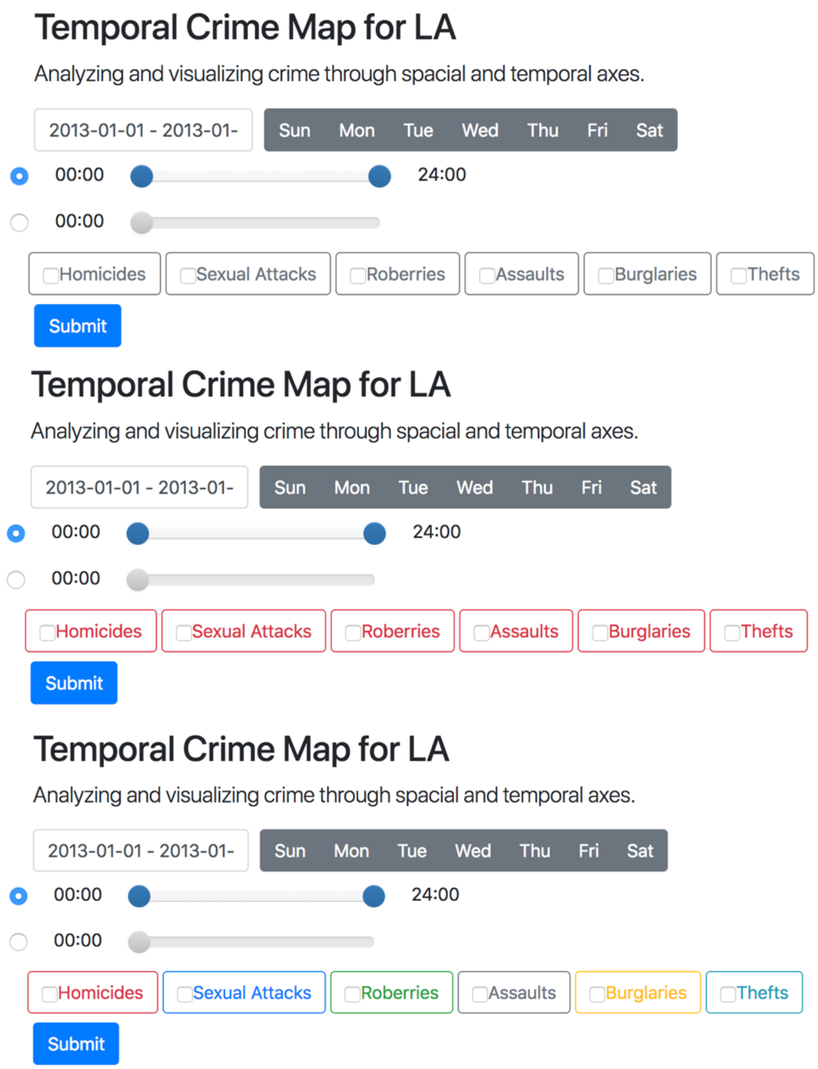
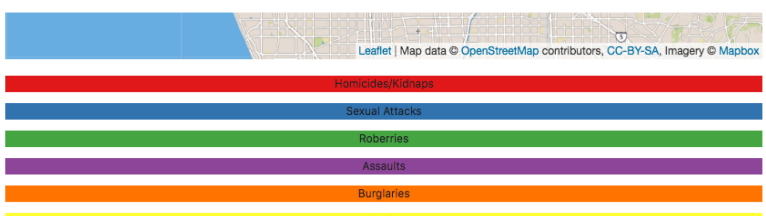


Figure 10. Checkbox Filtering Experimented Colors

Since we have not aligned our button colors corresponding to our criminal categories, we need a legend below our map to tell users which color stands for which class. In this experiment, we tried two different legends which are color bars in perpendicular and parallel to each other. As figures below, we can see legends in parallel can reduce the length of our whole map and more beautiful.



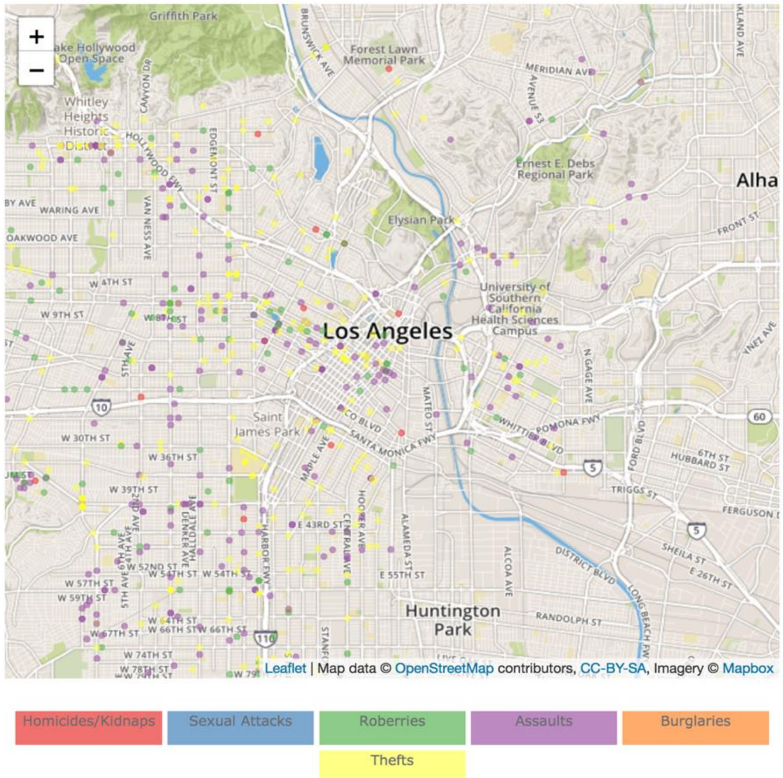


Figure 11. Different legend method below the color map

B. Visualization Result:

In the end, we merge buttons together below all buttons, change button colors to map the criminal types and remove legend blow the map. This figure is our final result. We can easily see now our users can use only one button to filter out the results they are not interested in. Moreover, using colored buttons can help them catch the criminal type more easily on the map.

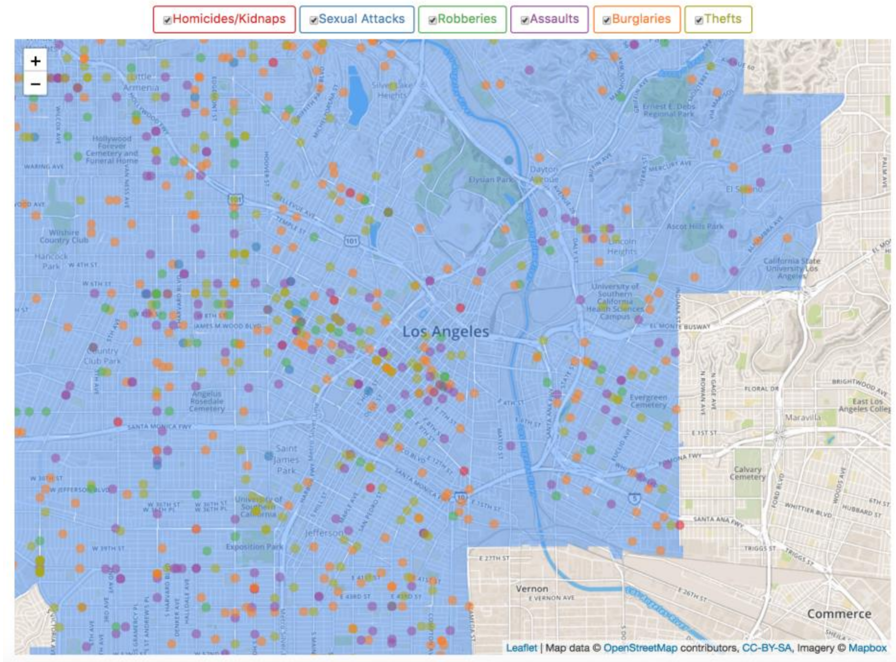


Figure 12. Checkbox Demo

C. Crime Prediction Model Training/Testing Experiments:

For the models, we tried Linear Regression and Gradient Boosting Models with a *window_length* of 90 and 180 days along with a *label_window_length* of a week (7) and a month (30).

Gradient Boosting Classifier:
Window_length, l*abel_window_length*: 180, 30
Running time: 38 minutes
R2 scores: 0.72/0.73

Linear Regression:

window_length / label_window_length	R2 scores for training/testing set	Execution Time (secs)
90 / 7	0.39 / 0.42	9
180 / 7	0.41 / 0.44	15
90 / 30	0.69 / 0.71	13
180 / 30	0.72 / 0.74	17

Table 3. Linear Regression prediction model testing results

VI. Evaluation

A. Prediction

After many experiments, we observed that using a more complicated Gradient Boosting Classifier does not yield better results. In fact, it performs a little bit worse than simple linear regression under the same window lengths. Furthermore, linear regression has much less running time (17 seconds vs. 38 minutes). So we decided to use a simple linear regression model rather than other complicated machine learning models.
Since this is a regression problem rather than a classification problem, it is natural to think of using R^2 score as the metric of evaluation. The higher the score, the higher relativity between the predicted results and the labels. And we can see that the R^2 score is more relevant to the *label_window_length*, which makes sense because it is harder to predict number of occurrences within a short window than a longer one. And longer window length does yield better results slightly, but while considering the computation time for each prediction on the server, we decided that it is appropriate to use a 90-day window.

B. **Method:**

We conducted an open ended survey. We randomly sampled 15 people on the Georgia Tech campus and on the Internet, showed them the baseline website along with our app's website, let them play with it, and asked them the following questions. (The Baseline website: https://www.trulia.com/real_estate/Los_Angeles-California/crime/)

C. **Evaluation Questions:**

1. What do you think about our website versus the baseline website?
2. What are the pros and cons of our approach?

D. **Results:**

Based on our survey results, which we conducted on randomly selected people on campus, we found that some people preferred more detailed information, while others did not. But all said that giving more information would give them the option to choose. While too many colors might be confusing, people found that it could be hard to distinguish one type of crime from another with only single colors. Therefore, we decided to make our app with more colors and to keep the timeline bar.

	Our App	Baseline websites
Pros	More detail Can select different type	Shows numbers of crime counts on map Simple design with one color
Cons	Many colors might confuse Too detail	No hour information

Table 4. Human evaluation results

VII. Schedule of Activities

The following were the plan of activities:

1. Create Flask web server
2. Plot crime points on map of LA
3. Filter crime points by crime type (midterm)
4. Visualize temporal aspects
5. Predict danger zones on map
6. Make recommendations to visitors of LA
7. Create graphs of crime statistics
8. Host on Google Cloud (final)

Activity 3 served as the midterm checkpoint, while Activity 9 served as the final checkpoint.

VIII. Technologies

Tables 2 summarizes the technologies used.

Technology	**Purpose**
Python Flask	Web Micro-framework
Leaflet.js	Mapping framework
D3, Google Data Studio	Frontend Charting
Scikit-Learn	Machine Learning
Google Cloud Platform	Hosting Website

Table 5. Table of technologies used.

IX. Team Contributions

All team members have contributed a similar amount of effort.

Table 3 shows the division of labor.

Member Name	**Responsibilities**
ShuHo Chou	Model Prediction
Kairi Kozuma	Frontend Visualization, Time Filtering, Flask App

Dennis Sosa	Color Scale Map, Same Data Point Crimes Feature
Chiamin Wu	Check Box Filtering
ChiLin Wu	Data Cleaning, Host App on Google Cloud

Table 6. Division of labor.

X. Conclusion

The constructed crime map successfully completed its objective of visualizing LA crime data. In conventional features of zooming and panning to locations of interest, our crime map offered advanced temporal filtering. This let users select the date range, time range, and the day of week to suit their needs. Furthermore, our crime map categorizes crime into six major types: homicides/kidnaps, sexual attacks, robberies, assaults, burglaries, and thefts. Each category has a distinct color, such that it is easy to glance at the map and recognize which crime types are widespread. Additionally, users are able to click on a point of interest to read small descriptions regarding the crime for the particular point on the map. Furthermore, the prediction model suggests areas of danger based on previous crime incidents. Based on our survey results, users were able to clearly distinguish the advantages of using our crime data tool versus modern crime data tools.

XI. References

[1] Almanie, T., Mirza, R., & Lor, E. (2015). Crime prediction based on crime types and using spatial and temporal criminal hotspots. *arXiv preprint arXiv:1508.02050*. https://arxiv.org/pdf/1508.02050

[2] Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. (2017). Crime prediction through urban metrics and statistical learning. *arXiv preprint arXiv:1712.03834*. https://arxiv.org/pdf/1712.03834

[3] Brantingham, P. L., Brantingham, P. J., Song, J., & Spicer, V. (2016). Crime Pattern Visualization. http://summit.sfu.ca/system/files/iritems1/16510/ICURS_Crime-Pattern-Visualization-2016.pdf

[4] C. Chang. In Los Angeles, homicides are down, but violent crime is up for the fourth year in a row. *Los Angeles Times*, 2017. http://www.latimes.com/local/lanow/la-me-lapd-crime-stats-20171230-story.html

[5] Eck, J., Chainey, S., Cameron, J., & Wilson, R. (2005). Mapping crime: Understanding hotspots. http://discovery.ucl.ac.uk/11291/1/11291.pdf

[6] Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). Visualizing criminal networks reconstructed from mobile phone records. *arXiv preprint arXiv:1407.2837*. https://arxiv.org/abs/1407.2837

[7] Hagenauer, J., Helbich, M., & Leitner, M. (2011, July). Visualization of crime trajectories with self-organizing maps: a case study on evaluating the impact of hurricanes on spatio-temporal crime hotspots. In *Proceedings of the 25th conference of the International Cartographic Association, Paris*. http://www.geog.uni-heidelberg.de/md/chemgeo/geog/gis/hagenauer_icc2011.pdf

[8] K. O'Connell and S. M. Staudaher and A. Godwin. Making Public Safety Data Accessible in the Westside Atlanta Data Dashboard. *Bloomberg Data For Good Exchange*, 2016. https://arxiv.org/abs/1609.09756

[9] L. J. S. Silva and S. F. Gonzalez and C. F. P. Almeida and S. D. J. Barbosa. CrimeVis:

An Interactive Visualization System for Analyzing Crime Data in the State of Rio de Janeiro . *International Conference on Enterprise Information Systems*, 2017. http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006258701930200

[10] Maltz, M. D. (2010). Look before you analyze: Visualizing data in criminal justice. In *Handbook of quantitative criminology* (pp. 25-52). Springer, New York, NY. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.169.5111&rep=rep1&type=pdf

[11] Rasheed, A., & Wiil, U. K. (2014, August). PEVNET: A framework for visualization of criminal networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (pp. 876-881). IEEE. http://ieeexplore.ieee.org/document/6921689/?part=1

[12] Sarvari, H., Abozinadah, E., Mbaziira, A., & Mccoy, D. (2014, May). Constructing and analyzing criminal networks. In *Security and Privacy Workshops (SPW), 2014 IEEE* (pp. 84-91). IEEE. http://www.ieee-security.org/TC/SPW2014/papers/5103a084.PDF

[13] Shah, S., Khalique, V., Saddar, S., & Mahoto, N. A. (2017). A Framework for Visual Representation of Crime Information. *Indian Journal of Science and Technology*, *10*(40). http://www.indjst.org/index.php/indjst/article/viewFile/120079/82703

[14] Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, *30*(9), 1694-1716. http://www.tandfonline.com/doi/full/10.1080/13658816.2015.1099658

[15] T. Dasgupta and A. Naskar. CrimeProfiler: Crime Information Extraction and Visualization from News Media. *Web Intelligence*, 2017. https://dl.acm.org/citation.cfm?doid=3106426.3106476

[16] Wang, B., Yin, P., Bertozzi, A. L., Brantingham, P. J., Osher, S. J., & Xin, J. (2017). Deep Learning for Real-Time Crime Forecasting and its Ternarization. *arXiv preprint arXiv:1711.08833*. https://arxiv.org/pdf/1711.08833