CSE 6242 Final Report - Team GTD
Members: Trishla Chokshi, Scott Goldstein, Hadrien Lacroix, Ben Javani, Jamie Zhong, Julia Zhu

Introduction – Motivation

Terrorism can be defined as "criminal acts, including against civilians, committed with […] the purpose to provoke a state of terror in the general public"[1]. It is a global issue evolving on a daily basis. This project builds on the Global Terrorism Database (GTD), which lists terrorist attacks from 1970 to 2018. It counts about 200,000 rows and 135 features detailing the type of attack, location, weapons used, motive, casualties, press coverage, etc. This project aims to help the general population explore the nature of terrorism, help researchers in their work, and provide additional tools to government agents focused on fighting terrorism. Additionally, discussions around terrorism tend to be emotionally charged. There is a disconnect between the threat reality, the media coverage, and the threat perception by the general public. Lafree[2] lists several misconceptions dispelled by rigorous analysis: Crenshaw[3] debunked the US State Department claim that one-third of all terrorist attacks worldwide are directed at the US, and Neumayer and Plümper[4] showed that most foreign victims of terrorist attacks are not US citizens. We hope that providing a tool that considers the aggregated reality of terrorism over 50 years rather than one isolated event, 200,000 data points instead of a unique one, built on facts only, and free of rhetorical or political motivations, will help people be more informed and less vulnerable to the fear-driving objectives of terrorists.

Problem definition

Users could more effectively understand trends in terrorism or combat the effects of terrorism if there was a better way to visualize and mine the data. Most of the work currently published around terrorism rarely produces empirical evidence. Lum et al.[5] reviewed 14,000 terrorism articles: only 3% used quantitative analysis. Silke[6] found that 80% of the literature is not research-based, but instead too narrative, condemnatory, and prescriptive. Some academic papers use statistical analysis, but visualization is limited to EDA and not interactive. This format is abstruse to the general public. While the GTD successfully collects information on terrorist events, it falls short in the user experience and its ability to represent the data in an interesting and visually appealing manner: the few existing interactive visualizations are outdated[7] or limited in interactivity[8]. Common query results include simple visualizations such as bar/pie charts and line graph or a list of results. Literature on terrorism focuses more on specific topics and less on creating open source tools to help consumers better explore the data.

Below, we outline our method to transform the GTD from a relational to a graph database, which will improve speed and performance, use machine learning to identify event features most linked to death count, and create an interactive map that allows users to interact with terrorism data for their specific needs.

Survey

Our survey explored four areas: data engineering, data pre-processing, analysis and visualization literature will be covered in the corresponding sections. On top of these four topics, communication and media coverage are worth studying when communicating about terrorism. It is paramount to frame our discourse and focus on topics that matter the most to the general public or on which there are the greatest misconceptions. Ritchie et al. [9] showed that terrorism is over-represented relative to its share of deaths in media coverage, and that the fact that the perpetrator is Muslim and was arrested contributes more to the coverage than the number of targets or casualties. Sunstein[10] demonstrated that people are strongly driven by emotion and fear when it comes to making decisions about how to respond to terrorism, which in turn has an impact on expenses, civil rights, and national security expenses (increased airport security, for example). Sunstein concludes that educating the general public will alleviate fear. Allouche and Lind[11] analyzed quantitative surveys of British and American citizens to better understand their perception of individual and nation-wide risk, counter-terrorism measures, reflex towards Muslim communities, etc. One caveat of the survey is its focus on Anglo-Saxon culture, so we would ideally complement it with a more global study.

Proposed method

Our approach develops in several steps. First, we select the features that were relevant to our analysis. Second, we convert the tabular database to a graph database to facilitate building a network graph and speed up analysis. Third, we start building the graph and exploring connections between individuals, terrorist cells, events, countries, weapons, etc. Fourth, we develop machine learning algorithms to predict total fatalities. We finally develop a website serving interactive visualizations presenting our work and enabling people to interact with it and answer their own questions.

Intuition - why should it be better than the state of the art?

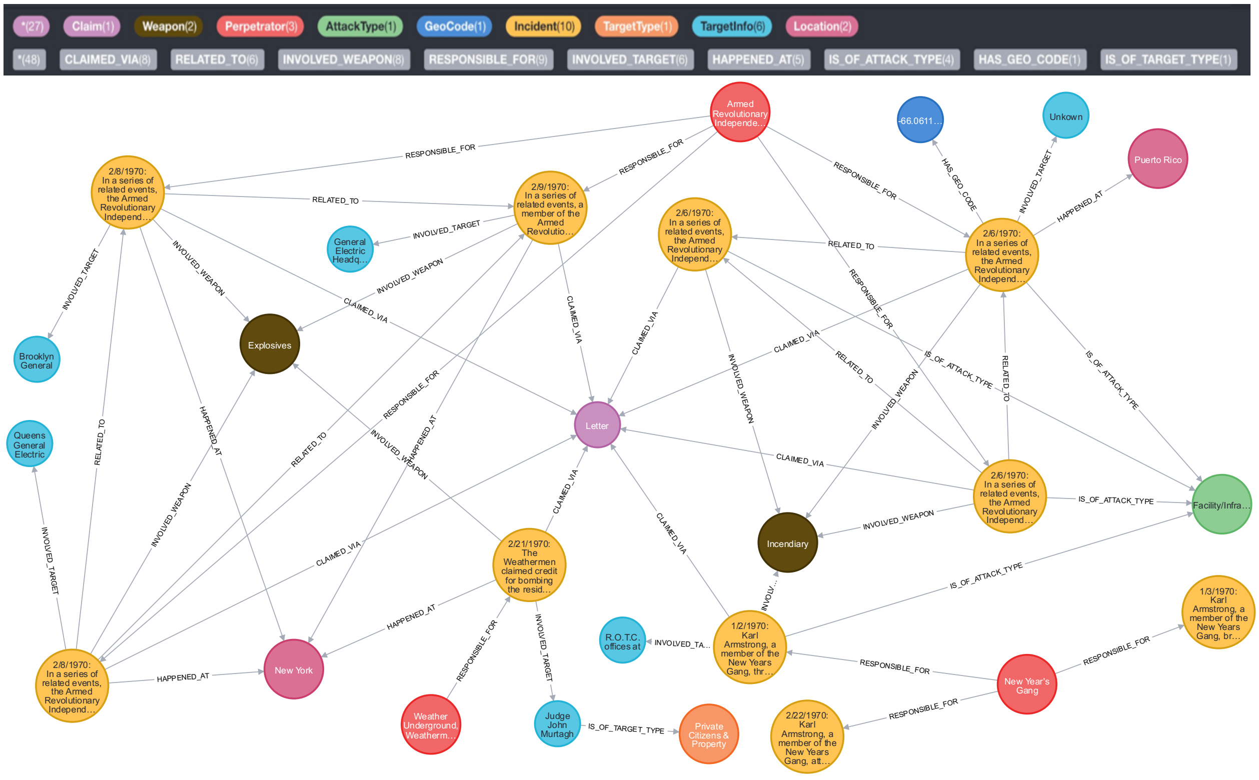
Our goal is to build interactive and accessible network graphs and visualizations using the Global Terrorism Database (GTD) to enable citizens, law enforcers, researchers, and policy makers to explore terrorism trends, impact, and distribution. The advantage over what is currently state of the art resides in:

- the *ease of access and understanding:* a URL serving modern interactive visualizations with the ability to filter on certain attributes.
- the *approach:* transforming the tabular dataset into a network dataset and building the first network graph based on the GTD. Having all entities and attributes of a given terrorism incident in one row is not only a very inefficient design from a database design perspective, but also fails to explicitly show the relationships between entities.

Description of your approaches: algorithms, user interfaces, etc.

Data engineering

These short comings motivated our team to design a graph model based on GTD data and store the data in a Neo4j Database. We rely on the Neo4j documentation[12] for this step and write a Python script to read the GTD Excel file, build the nodes and relationships, and persist it to the graph database. We also created a Neo4j Bloom environment. This allows for enhanced visual data exploration and codeless querying of the dataset. An example query and it's resulting graph can be seen below:


*(27)
Claim(1)
Weapon(2)
Perpetrator(3)
AttackType(1)
GeoCode(1)
Incident(10)
TargetType(1)
TargetInfo(6)
Location(2)
*(48)
CLAIMED_VIA(8)
RELATED_TO(6)
INVOLVED_WEAPON(8)
RESPONSIBLE_FOR(9)
INVOLVED_TARGET(6)
HAPPENED_AT(5)
IS_OF_ATTACK_TYPE(4)
HAS_GEO_CODE(1)
IS_OF_TARGET_TYPE(1)
Armed Revolutionary Independe...
-66.0611...
Unkown
Puerto Rico
2/8/1970: In a series of related events, the Armed Revolutionary Independ...
2/9/1970: In a series of related events, a member of the Armed Revolutio...
General Electric Headq...
2/6/1970: In a series of related events, the Armed Revolutionary Independ...
2/6/1970: In a series of related events, the Armed Revolutionary Independ...
Brooklyn General
Explosives
Letter
Queens General Electric
2/6/1970: In a series of related events, the Armed Revolutionary Independ...
Facility/Infra...
Incendiary
2/21/1970: The Weathermen claimed credit for bombing the resid...
1/2/1970: Karl Armstrong, a member of the New Years Gang, thr...
R.O.T.C. offices at
1/3/1970: Karl Armstrong, a member of the New Years Gang, br...
2/8/1970: In a series of related events, the Armed Revolutionary Independ...
New York
Weather Underground, Weatherm...
Judge John Murtagh
Private Citizens & Property
2/22/1970: Karl Armstrong, a member of the New Years Gang, att...
New Year's Gang
RESPONSIBLE_FOR
RELATED_TO
INVOLVED_TARGET
INVOLVED_WEAPON
CLAIMED_VIA
HAPPENED_AT
HAS_GEO_CODE
IS_OF_ATTACK_TYPE
IS_OF_TARGET_TYPE

Data processing for machine learning

47% of the values in GTD are missing, and 74% of the features have missing values. Pagán[13] tried various imputation and discretization methods and quantified the impact on the error rate of several classification models. For example, they found that median imputation, where all missing values are replaced by the median of all known values for that attribute, produced the highest accuracy.

To process our data for analysis, our first step is to remove features with values missing for more than half of the observations. Next, we convert all features with text to categories and features with numerical categories to factors. To account for missing values, mean and median imputation are performed on the numerical data and mode imputation is performed on the categorical data.

Non-trivial analysis

Looking at influential factors is paramount to our goal of predicting future terrorist attacks and their potential fatalities. Guohui et al.[14] used correspondence analysis – a technique used to analyze two or more categorical variables as points in low dimensional space[15] – to identify dimensions with highest variable dispersion. They found economic development, explosive weapons, and military targets as the most important factors contributing to high fatality levels. Mo et al.[16] applied Logistic Regression algorithms to predict attack types, while Kim et al.[17] used deep learning based named entity recognition to identify potential threats and prevent terrorist attacks; however, these approaches have a significant disadvantage because they rely on a large labeled dataset for model training.

For our project, elastic net is used for variable selection in order to determine the most important factors affecting total number of fatalities for a given terrorist attack. Elastic net, being a combination of LASSO and ridge regressions, performs both variable selection and regularization and therefore combines the strengths and drawbacks of both. We test a variety of alphas and l1 ratios and select the combination that most reduces the mean squared error.

Once the most important features have been selected by our elastic net model, we create a regression model to predict fatalities. We create interaction plots and inspect the scatter plots and residual vs fitted plot of each variable to determine if our model requires interaction or nonlinear terms. We tested 4 different models: CART, Linear Regression, Adaptive boosting, and Neural Network. Intuitively, AdaBoost works by evaluating the performance of explanatory variables iteratively on samples of observations from the entire dataset, placing more emphasis on correctly classifying mis-classified observations from previous iterations. AdaBoost also has the advantage of having a faster computational time than other machine learning algorithms, making it ideal to implement for a large dataset like ours. On the other hand, a simple linear regression may be a good fit for our data. Therefore, we should use AdaBoost only if it provides a significant boost in accuracy over linear regression. Our two other classification models, neural network and CART, were also evaluated based on simplicity and their results through run-time and confusion matrices.

Graph database and interface

As graph databases store relationship information as a first-class entity, graph databases are very adept at working with evolving relationship networks. In addition, the flexibility of a graph database model allows for the addition of new nodes and relationships without compromising the existing network or expensively migrating data. With data relationships at their center, graph databases are highly efficient when it comes to query performance, even for deep and complex queries.

Our team's approach for designing the data model (Figure 1) was to divide the GTD data column fields into entities and attributes, and different relationships were introduced to connect the entity nodes in the graph. There are 120 separate attributes of each incident under multiple broad categories. In order to design the graph nodes, we used a heuristic approach and created the nodes and the relationships between them, few of them shown below and the rest in the Appendix:

1. *Year:* specifies the year of the incident.

⋮

13. *Region:* a 12-category region where each incident country belongs to.

User Interface Functionality

Previous researchers have used the GTD to describe the spatiotemporal trends[18,19] and used hybrid methods that incorporated multiple panels on the same page, comprising of a choroplethic map for geographic analysis, a slider for temporal analysis, and panels to select activities[20]. We can build upon their work and improve it with greater interactivity, for example, by incorporating tooltips into the map view for at-a-glance statistics about relevant countries and by adding our nontrivial graph algorithms. Salem and Naoali[21] used pattern recognition to discover clusters in the GTD. They acknowledged a gap was not performing clustering on the quantitative features (something we aim to do).

The frontend uses a combination of VueJS and D3 to present the incident and build the interactions. The data is accessed through a REST API hosted on an express server. Users will be presented with a choropleth map. The countries will be color coded according to the number of casualties by terrorism that have occurred in that country; darker colors will indicate more casualties (Figure 3). When users query a country and date range, the map will show only the top 50 queried incidents (Figure 4). Incidents will be represented as nodes on the choropleth map with higher casualty incidents having larger nodes. Users can also get a closer look by zooming in (click or pinch to zoom; drag the map to pan). Hovering over an incident will show more details (Figure 5). These details include 7 features chosen by our Elastic Net algorithm. In the far-right column, users can see the actual number of casualties and our Machine Learning algorithms prediction.

Users may also upload additional nodes in JSON format. Nodes do not need to be incidents, but the data should include styling information – e.g. the label, size, color, and/or position of the node. This feature enables users to explore new correlations and perform custom analysis with data outside of the GTD. To demonstrate this feature, the application will include casualty predictions from the machine learning algorithm described previously. These nodes will be sized according to the number of predicted casualties and then positioned to overlay the terrorist incident being labeled. This will give a visual representation of the algorithm's accuracy; better accuracy results in prediction nodes sized similarly to incident nodes.

Application Architecture

The visualizations will be hosted on the internet using a Model-View-View Model (MVVM) architecture. The Model is a Neo4J instance hosting GTD data; the View is a VueJS application that utilizes D3 for building a choropleth map; and the View Model is the representation of the data as visual components on the page. A GraphQL API Layer will act as a median between the Model and the View. This well-established pattern allows the view to be developed independently from the domain and accounts for differences in the View Model and the Model (Syromiatnikov). The application will be hosted on a single EC2 instance in an AWS Cloud. The instance will publicly expose an endpoint that serves the application.

Experiments/ Evaluation

Machine learning predictions

We will evaluate our machine learning models by calculating mean squared error between the outputs and the values in the original GTD. We will also calculate the R^2 of each model. The total fatalities variable will be converted into a categorical variable with the same percent of data in each category. For example, 0-10 fatalities could be a category. Then, a confusion matrix will be used to evaluate our output.

Web application experiments

We will ask users to find how many casualties were the result of terrorism and how many casualties were a result of the September 11 attacks using our application. We will monitor the users as they perform these tasks and ask the users to fill out a survey once they have completed the tasks. Survey questions include: How likely are you to explore the website? How engaging is the design (colors, structure) of the website? Does the website appear easy to navigate? How well does the website convey terrorism data? How likely are you to recommend this website to someone interested in learning more about terrorism trends? What improvements would you make to the design of the website?

Details of the experiments/observations

Web application

We asked family members to perform the web application experiment so that we could directly observe their behavior. When tasked with finding how many casualties were the result of terrorism, all attempted to find the answer through querying the database even though the total was in the title of the page. When querying for information, none of the users used the title feature. After querying to get incidents, some users struggled to find the correct incident because nodes were not labeled with city names on the map. However, all users were able to find the correct results by the end of the experiment without major difficulties.

Some of the more telling results from the survey can be found in the Appendix. Overall, we found that the users really liked the design, including color and structure. However, we found that users who might not be as experienced with databases or queries, found the website slightly difficult to navigate.

Machine learning

We use elastic net to select the 7 most important features for predicting the number of kills out of the over 100 total variables. The results of the elastic net model can be found in Figure 1 in our Appendix. Tuning various values for alpha, we find that the value that yields the best results is alpha = 0.1 (refer to Figure 2 in appendix). The coefficient plot shows the relative importance of each feature, from which we conclude the number wounded, suicide, weapon type, success, target type, weapon subtype, and terrorism certainty are the most important.

Our final model and the results of the rest of the models (confusion matrices) can be found in the appendix. Although we found the MSE to be lowest for the classification tree, at 72.54, this is not significantly smaller than the MSE of the linear regression, at 76.26. Classification has the drawback of requiring us to bin our data into several categories, which loses out on some granularity. In light of this, as our classification models did not provide a significant advantage over the simple linear regression model, we use this model for prediction. Our R^2 was 0.375, but given that we are attempting to analyze human behaviors, this is expected—R^2 values of below 0.5 are common due to the heterogeneity in human behaviors.

Our final predictive model is as follows:

$$\begin{aligned} nkill = 0.11 + 0.12nwound + 6.7suicide + 0.35weapon\_type + 1.52success \\ + 0.18terrorism\_certainty + 0.03target\_type - 0.18weapon\_subtype \end{aligned}$$

Graph interface

The figure in the Appendix shows is an example graph visualization showing seven random incidents claimed via letter, with the legend on top.

Conclusions and discussion

The graph database has allowed us to very quicky query incidents based on any feature of our choice, weapon or terrorist group, for example, and then also go a step deeper and analyze related incidents.

The map visualizations enable users to learn about terrorism at different levels. At a high level, users can see which regions of the globe have been most impacted by terrorism through colors on the choropleth map. Through the querying feature, users can then see which regions within countries are predominantly affected by terrorism through the clustering of nodes. Hovering over nodes gives users granular details about incidents and which factors are most important in determining the number of casualties. Finally, users can compare results from multiple queries to get a sense of how terrorism may differ between regions. While there are improvements to be made to the usability of the interface, we believe that the visualizations and interactions of the application make insights into terrorism trends more accessible.

On the machine learning side, our objective was to predict the number of casualties for the 5.76% of observations missing this data. We used an elastic net model to select the most important features for prediction and tested the performance of four separate models and chose the linear regression model for prediction. On the test set, this model yielded an R^2 of 0.375 and an MSE of 76.26.

Overall, compared to previous studies performed on the GTD, our project has significantly increased ease of access to the data, both through our front-end map visualizations and graph database, and our machine learning models have provided techniques for predicting unknown data values and determining important features. Future studies can build upon our work by adding to our visualization and providing more charts/graphs for users to explore. They can also perform graph algorithms on our graph database.

Distribution of team member effort

All team members contributed equally, Ben processed the data and built the graph database. Scott, Trishla and Julia focused on the machine learning predictions. Jaime designed the front end. Hadrien and Trishla wrote the report.

[1] United Nations Security Council. (2004). Resolution 1566. https://www.un.org/ruleoflaw/files/n0454282.pdf

[2] LaFree, G., Laura, D., Miller, E. (2015). Tracking worldwide terrorism trends. Putting Terrorism in Context, Lessons from the Global Terrorism Database. p. 17-49. New York: Routledge.

[3] Crenshaw, M. (2001). Why America? The Globalization of Civil War. Current History 100 (December). p.425-432.

[4] Neumayer , E., Plümper, T. (2011). Foreign Terror on Americans. Journal of Peace Research 48. p.3-17. https://doi.org/10.1525/curh.2001.100.650.425

[5] Lum, C., Kennedy, L. and Sherley, A. (2006). Are Counter-Terrorism Strategies Effective? The Results of the Campbell Systematic Review on Counter-Terrorism Evaluation Research. Journal of Experimental Criminology 2:489–516. DOI: 10.1002/14651858

[6] Silke, A. (2004). The Road Less Travelled: Trends in Terrorism Research. Research on Terrorism: Trends, Achievements and Failures. p.186-213. Routledge.

[7] University of Maryland. (n.d.). GTD Explorer. http://www.cs.umd.edu/hcil/gtd/gtd/explorer.html

[8] Tilburg University. (n.d.). Map of all the attacks between 2011 and 2014. Global Terrorist Attacks. https://public.tableau.com/views/InteractiveVisualization_0/Dashboard3?:display_count=yes&:embed=y&:showTabs=y&:showVizHome=no

[9] Ritchie, H., Hasell, J., Appel, C., Roser, M. (2019) Terrorism. Our World In Data. https://ourworldindata.org/terrorism

[10] Sunstein, C. R. (2003). Terrorism and Probability Neglect. Journal of Risk and Uncertainty. https://doi.org/10.1023/A:1024111006336

[11] Allouche J.; Lind J. (2010) “Public attitudes to terrorism” Public Attitudes to Global Uncertainties. UKRI. https://esrc.ukri.org/files/public-engagement/public-dialogues/full-report-public-attitudes-to-global-uncertainties/

[12] https://neo4j.com/docs/

[13] Pagán J.V. (2010). Improving the classification of terrorist attacks a study on data pre-processing for mining the Global Terrorism Database. 2nd International Conference on Software Technology and Engineering. pp. V1-104-V1-110, https://doi.org/10.1109/ICSTE.2010.5608902

[14] Guohui, L., Song, L., Xudong, C., Hui, Y., & Heping, Z. (2014). Study on Correlation Factors that Influene Terrorist Attack Fatalities Using Global Terrorism Database. Procedia Engineering, 84, 698-707. https://doi.org/10.1016/j.proeng.2014.10.475

[15] Clausen, S. E. (1998) Applied correspondence analysis: an introduction. Sage. https://us.sagepub.com/enus/nam/book/applied-correspondence-analysis

[16] Mo, H., Meng, X., Li, J., Zhao, S. (2017) “Terrorist Event Prediction Based on Revealing Data,” International Conference on Big Data, 239-244. https://doi.org/10.1109/ICBDA.2017.8078815

[17] Kim, I., Pottenger, W. M., and Behe, V. (2018) “Can a Student Outperform a Teacher? Deep Learning-based Named Entity Recognition using Automatic Labeling of the Global Terrorism” IEEE International Symposium on Technologies for Homeland Security. p. 1-6. https://doi.org/10.1109/THS.2018.8574179

[18] Wang, X., Miller, E., Smarick, K., Ribarsky, W., & Chang, R. (2008). Investigative Visual Analysis of Global Terrorism. Computer Graphics Forum, 27 (3), 919–926. https://doi.org/10.1111/j.1467-8659.2008.01225.x

[19] Huamaní E., Alicia A., Roman-Gonzalez A. (2020). Machine Learning Techniques to Visualize and Predict Terrorist Attacks Worldwide using the Global Terrorism Database. International Journal of Advanced Computer Science and Applications (IJACSA). 11(4). http://dx.doi.org/10.14569/IJACSA.2020.0110474

[20] Webb, J., Cutter, S. L. (2009). "The Geography of U.S. Terrorist Incidents, 1970-2004" Terrorism & Political Violence. p.428-449 https://doi.org/10.1080/09546550902950308

[21] Ben Salem, S., Naoali, S. (2016). Pattern Recognition Approach in Multidimensional Databases: Application to the Global Terrorism Database. International Journal of Advanced Computer Science and Applications (IJACSA), 7(8). http://dx.doi.org/10.14569/IJACSA.2016.070838

Appendix

Graph Database

Complete list of nodes and relationships in our graph database:

1. *Year:* specifies the year of the incident.
2. *Incident:* includes ID, date, summary text, casualty, damage, etc.
3. *GeoCode:* the exact latitude and longitude of the incident's location.
4. *AttackType:* includes the type of the attack: assassination, kidnapping, bombing, etc.
5. *Claim:* whether a group claimed the responsibility of the attack and type of claim.
6. *Weapon:* describes the weapon used in attack, includes weapon name, type, sub type.
7. *Perpetrator:* describes perpetrator group or individual information.
8. *Hostage:* information on number of hostages, ransom demand, etc.
9. *TargetType:* the nature of the target such as government, military, police, business, etc.
10. *TargetInfo:* specifies the exact target name or entity.
11. *Location:* describes the attack place which includes city and state.
12. *Country:* describes the attack place which includes country name and country code.
13. *Region:* a 12-category region where each incident country belongs to.

Machine Learning

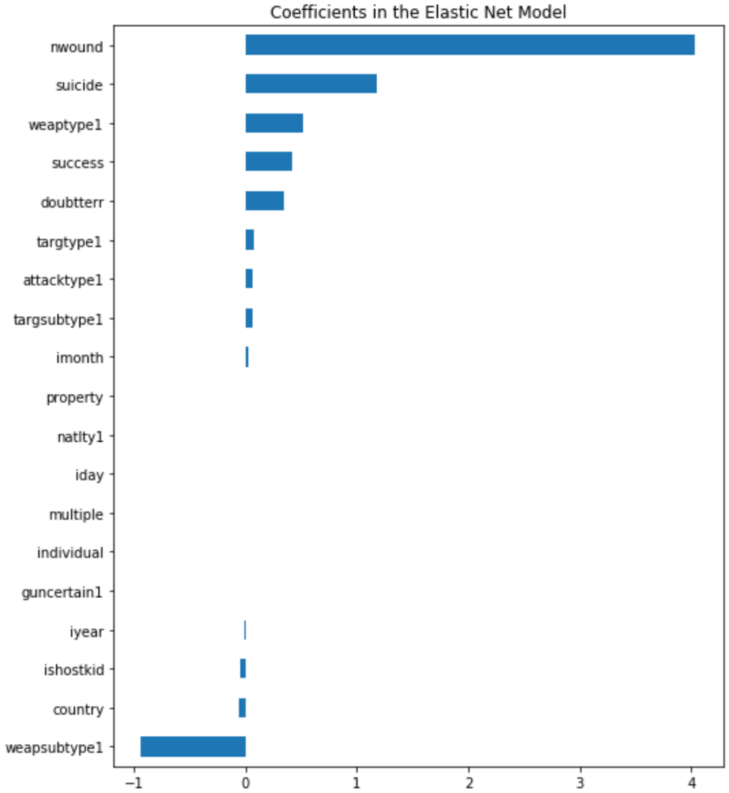


Figure 1: Ranking of features based on Elastic Net model

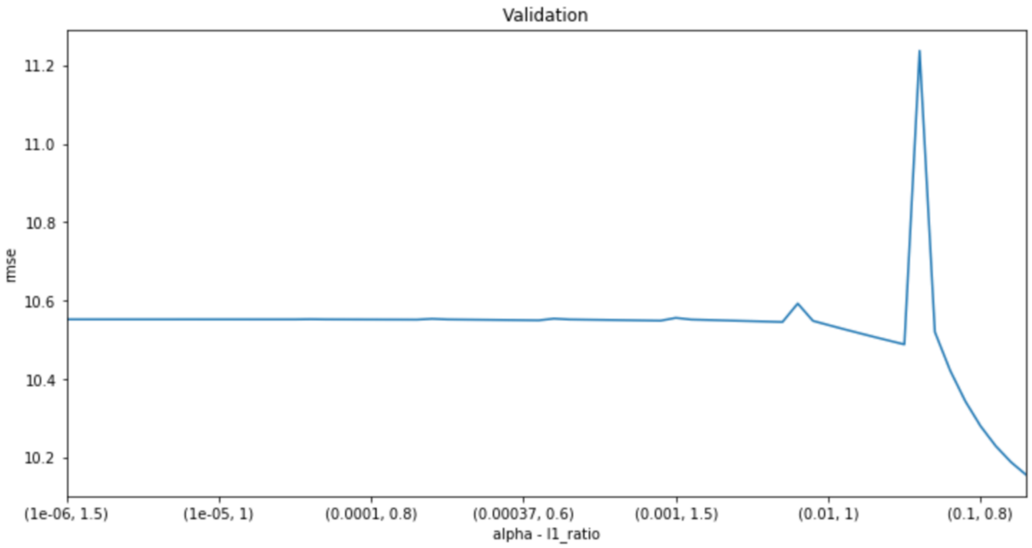


Figure 2: Alpha vs MSE graph for Elastic Net

Confusion matrices

AdaBoost

Predicted	***0***	***1***	***20***	***All***
Actual				
0	25592	0	2005	27597
1	9675	0	1946	11621
2	3348	1	1207	4556
20	4864	0	5431	10295
100	22	0	40	62
All	43501	1	10629	54131

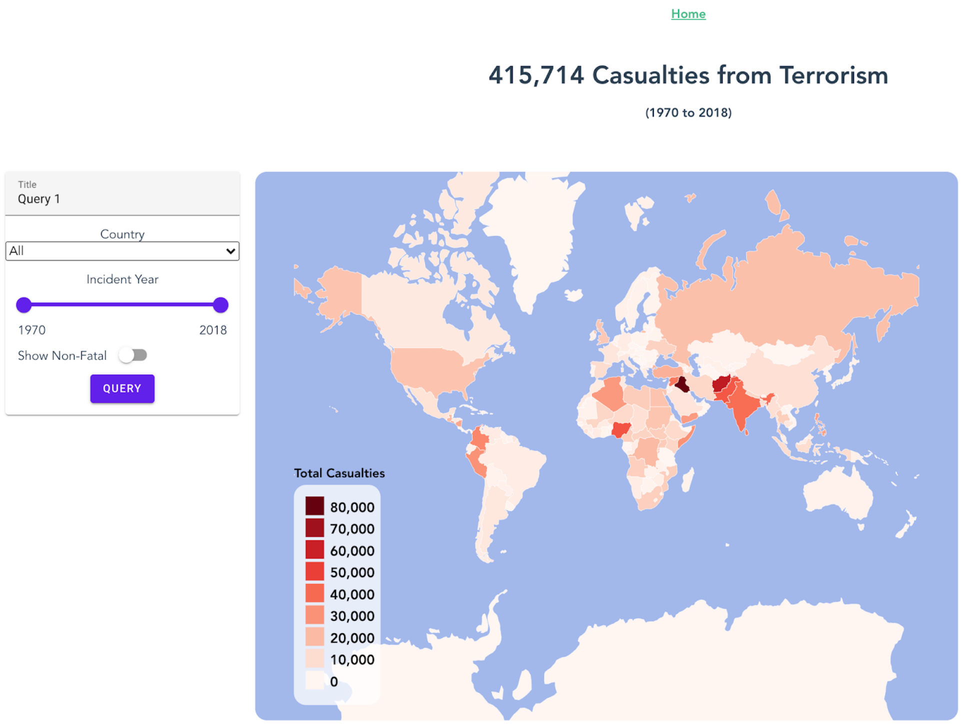
Neural Network

Predicted	***0***	***1***	***20***	***All***
Actual				
0	22441	4075	1081	27597
1	3692	6698	1231	11621
2	1516	2205	835	4556
20	2518	3540	4237	10295
100	17	16	29	62
All	30184	16534	7413	54131

CART

Predicted	***0***	***1***	***2***	***20***	***100***	***All***
Actual						
0	23942	2069	143	1442	1	27597
1	3963	5881	132	1645	0	11621
2	1695	1620	107	1134	0	4556
20	2695	1903	237	5444	16	10295
100	13	6	0	36	7	62
All	32308	11479	619	9701	24	54131

Web Application



415,714 Casualties from Terrorism

(1970 to 2018)

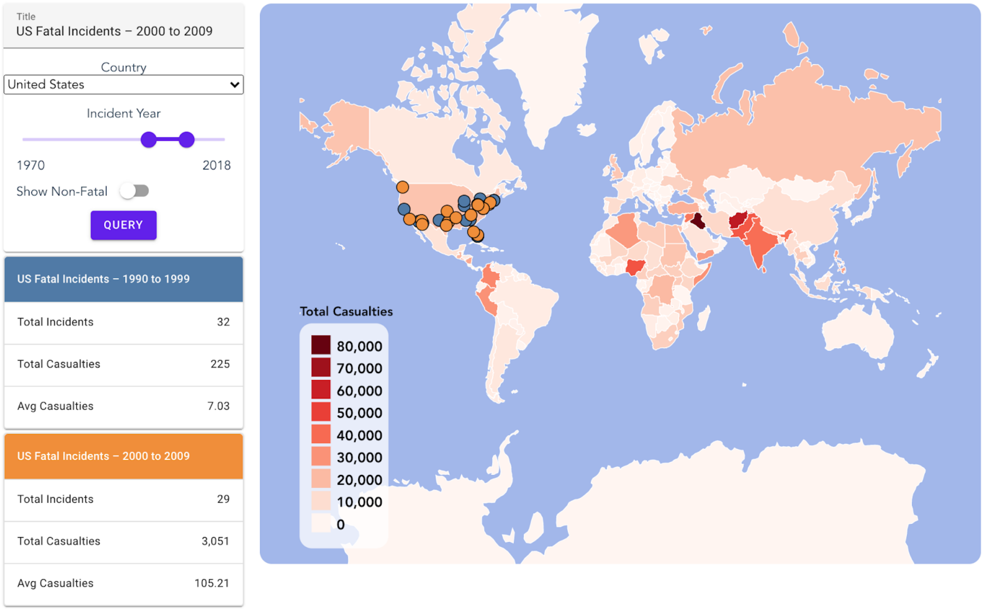
Title: Query 1

Country: All

Incident Year: 1970 – 2018

Show Non-Fatal

Figure 3: Website homepage



Title
US Fatal Incidents – 2000 to 2009

Country
United States

Incident Year
1970 2018

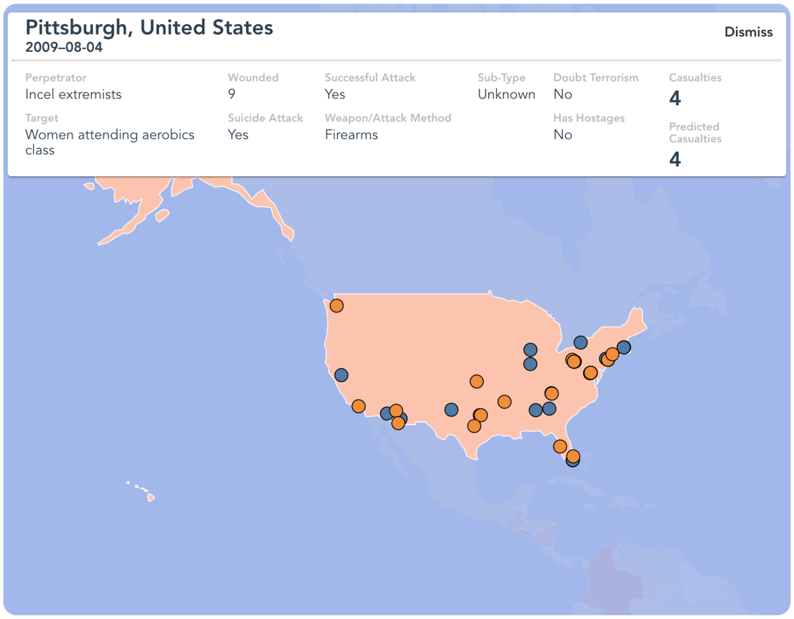
Show Non-Fatal

QUERY

US Fatal Incidents – 1990 to 1999	
Total Incidents	32
Total Casualties	225
Avg Casualties	7.03

US Fatal Incidents – 2000 to 2009	
Total Incidents	29
Total Casualties	3,051
Avg Casualties	105.21

Figure 4: Example queries based on country and year filter



Pittsburgh, United States
2009–08–04

Dismiss

Perpetrator: Incel extremists

Target: Women attending aerobics class

Wounded: 9

Suicide Attack: Yes

Successful Attack: Yes

Weapon/Attack Method: Firearms

Sub-Type: Unknown

Doubt Terrorism: No

Has Hostages: No

Casualties: 4

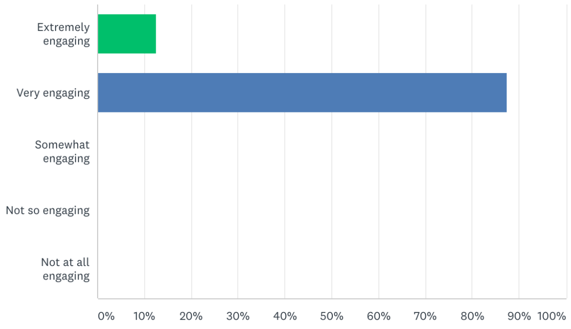
Predicted Casualties: 4

Figure 5: Zoom in to country with additional data displayed when hovering over datapoint

Web Design Survey

How engaging is the design (colors, structure) of the website?

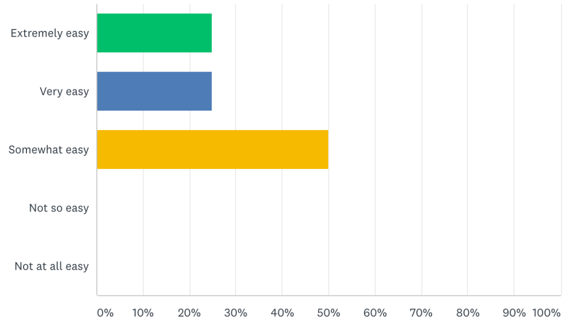
Answered: 8   Skipped: 0



ANSWER CHOICES	RESPONSES	
Extremely engaging	12.50%	1
Very engaging	87.50%	7
Somewhat engaging	0.00%	0
Not so engaging	0.00%	0
Not at all engaging	0.00%	0
TOTAL		8

Does the website appear easy to navigate?

Answered: 8   Skipped: 0



ANSWER CHOICES	RESPONSES	
Extremely easy	25.00%	2
Very easy	25.00%	2
Somewhat easy	50.00%	4
Not so easy	0.00%	0
Not at all easy	0.00%	0
TOTAL		8