World Terrorism Attack Analysis: A Data Science Approach

Aaron A. Gauthier, MBA, M.S.

May 7, 2020

1. Introduction

Many years ago, my life changed – many people's lives changed dramatically, and the world have never been the same for many us. We are all the summation of our life experiences and I am no different. We are all aware what happened on September 11^{th}, 2001 – however very few know how my life changed. I still remember that day vividly. I was sitting in a class at the University of Pittsburgh (PITT) in the Cathedral of Learning.

Unbeknownst to me as I was sitting in the class listening to the professor, my fellow classmates and chatting about what we had read over the weekend – September 11^{th} events were already in motion. At approximately 0846 the first commercial airliner hit the World Trade Center (WTC). At 0852 the University announced that all classes were cancelled for the day and campus was shutting down. We were all told to go back to our dorm rooms or apartments completely unaware of what just happened and what was about to happen in a rural field just outside of Pittsburgh, Pennsylvania. I welcomed the morning break – still no knowing what had just happened since I was in class. I made it back to my dorm room around 1030 in the morning and was alerted by one of my fellow colleagues in the Reserve Officers Training Corps (ROTC) about what had happened. I honestly was in disbelief – I thought it was some sort of strange joke being played on me. I turned on my TV and sure enough it was on the news. My life changed forever that day.

I hope through research I will be able to continue to serve community. I opted to focus on the application of classical machine learning to predict which terrorist groups an attack maybe attributable to based on the data from The Global Terrorism Database (GTD), National

Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland. (2019). The Global Terrorism Database™. The goal of this research is to address the looming terrorism threats, possible mitigations, and addressing the operationalization of trending disruptive technologies. Given the complexities and psychology involved in the research discipline of terrorism – I believe with my cross-discipline education and experiences I am uniquely qualified to contribute to such research. I hope to bridge the gap between theory and practice (research and operationalization).

The goal of this project was to start with the application of Classical Machine Learning (ML) and to evolve into applying Deep Learning (DL) frameworks including Natural Language Processing (NLP), Network Analysis, Game Theory, amongst others. The ultimate goal of my research is to create a recommendation engine based on available data to inform and shape senior decision makers thinking – a lofty goal. Starting with a classical ML approach was where I decided to start since I needed to establish a research baseline and evolving that baseline of research into the newest areas of technology and methodologies.

1.1. Terrorism Incident Analysis Challenges

- Multiple definitions of terrorism or terrorism incident (over 200)
- Prediction of future terrorism attack incidents
- Attributing unclaimed attack incidents to a group based on data attributes
- Bridging the gap between psychology opinion and evidential fact
- Utilizing data to inform strategic and operational decisions
- Informing policy through data

- Connecting the social sciences with the discipline of data science
- Numerous attack incidents to track
- Difficulty tracking attack incident dates and location
- Emerging and disruptive technologies
- Challenge identifying groups responsible for attack incidents
- Cultural, Political, Economic, Geo-Political and Media Bias related research
- Willingness of scholars to engage in respectful debate about common concerns
- Terrorism Policy interpreted from our perspective - not ground truth

2. Exploratory Data Analysis (EDA)

I downloaded the entire GTD extract which encompassed 135 features with 191, 465 observations. I chose to utilize Microsoft® Excel as my primary means of data cleansing activities. It is incredibly fast and efficient, and I was able to impute over ~8 million observation values in less than two hours. I utilized python for basic statistics, data frame and header information. I combined the separate 'iyear', 'imonth' and 'iday' features into a common date time-stamped 'Date' column - this increased the total feature count to 136 and total observations to 186, 463. I removed duplicate observations and incomplete observations which reduced my total observation count. I noticed and it is annotated in the GTD Codebook (October 2019) that there is a gap in the data. Data from 1993 is completely missing. This creates some anomalies in our data visualizations and models. A possible work around is to identify other terrorism databases that may have 1993 data and can be reliably matched up against GTD data to fix the gap in data during 1993. When producing the data visualizations with Tableau®, fourteen former countries like the Soviet Union, East Germany, Yugoslavia (just a few examples) and their associated data were removed due to change in the geo-political environment such as the former Czechoslovakia splitting into Slovakia and the Czech Republic.

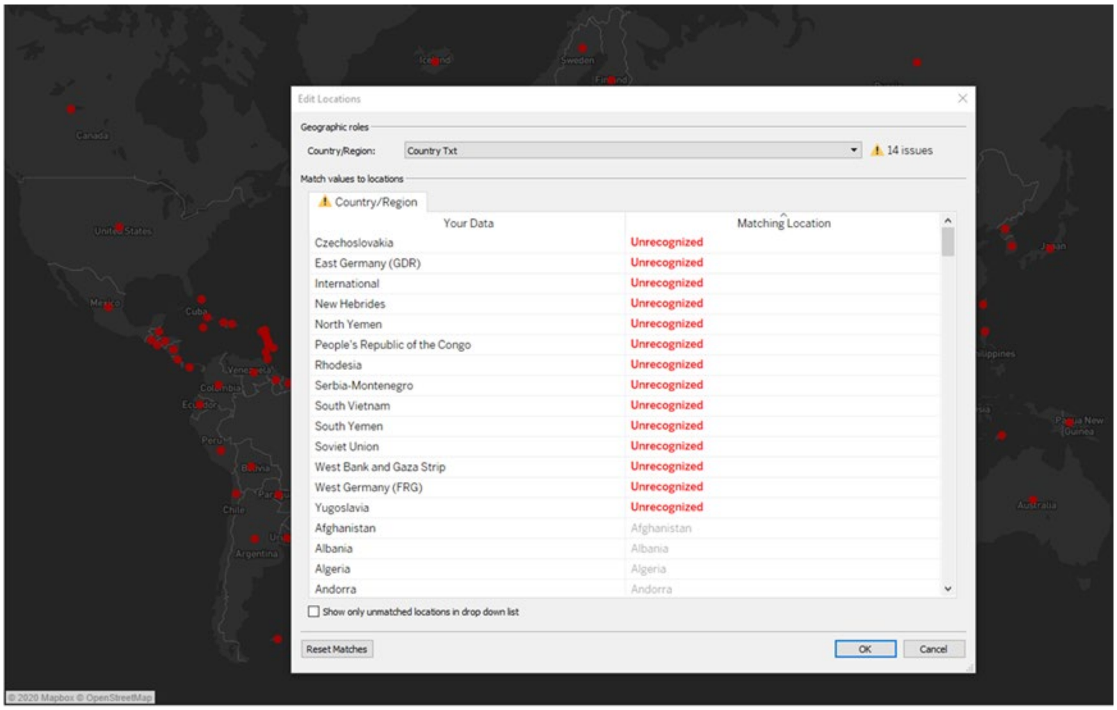

Figure 2: Former Countries Removed from Dataset

There are multiple other examples which when I have more time – I will figure out how to incorporate this data in the artifacts. Imputations were recorded, including new categorical variables however listing them out would take up pages of valuable research paper real estate. I plan on incorporating these additional categorical variables and imputation assumptions on the project website which is still a work in progress. I encourage the community to explore the GTD data and download the GTD Codebook to gain a more intimate understanding of the data and data collection methodologies through the years.

2.1. Classical Machine Learning (ML) Approach

Classical ML was chosen first to establish an ML baseline prior to exploring DL and other advanced methodologies. I utilized three models with two distinct feature lists with the target remaining common – Group Names. I utilized the Classification and Regression Trees (CART), K-Nearest Neighbors (KNN) and Random Forest (RF) models from SciKitLearn. In each of the three models, two different feature lists were utilized – one list with 18 features and another list with 60 features out of the total 136 features available. Hyperparameters were tuned utilizing randomized search.

3. Data Visualization and Analysis

I utilized Tableau® primarily to create the data visualizations and dashboards. Creating 105 data visualizations would have taken months and was not feasible to utilize in the timeframe for this project. Tableau® allowed me to create visualizations that can be embedded into a HTML website quickly and easily. It also allowed me to do some EDA on the data allowing me to find former countries that no longer exist (***see Figure 2 above***). Below are some good examples of the high-quality data visualizations that I created with Tableau® (***see Figures 3 & 4 below***).

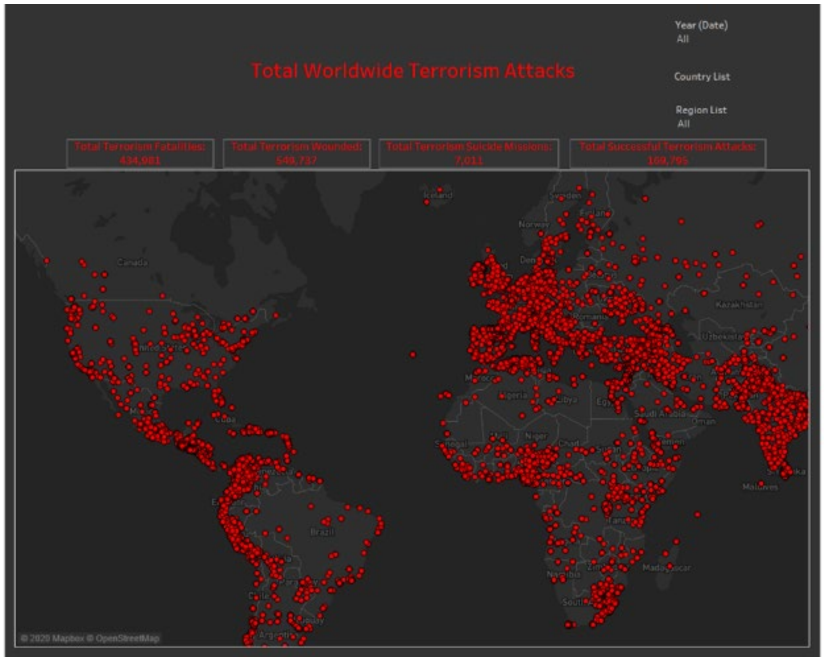

Figure 3: Total World Terrorism Attacks with Counters

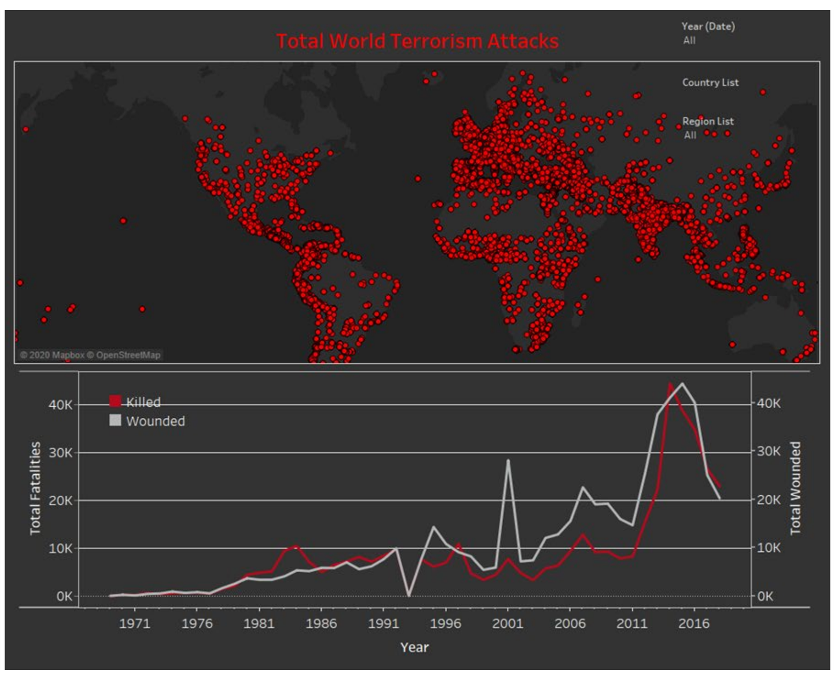

Figure 4: Total World Terrorism Attacks Dashboard with Line Plots

3.1. Country Borders Change Over Time

It was interesting to run across within Tableau® when I was plotting data on maps like in the above dashboards (**Figures 3 & 4 above**) that some countries no longer existed (**refer to Figure 2 above**). They had since to exist due to world events such as the fall of the Soviet Union, Berlin Wall, as well as other reasons. I know I was not thinking about this and was reminded about it thanks to my choice of using Tableau® to visualize data. This suggests an interesting challenge of how to deal with this data. Do we incorporate it into our models and visualizations? Is it historical data – or do we place the data where it lies and attribute it to the new country? Historically this does not make sense since it did not exist prior the current new country. How do we account for it in our models?

3.2. Terrorism Groups Change Over Time

Not only do country borders change over time, but so do terrorism groups. Terrorism Groups fracture combine and morph throughout their lifecycles. How do these changes affect our historical analysis or future predictions? Do we need to map these terrorist organizations back to their original state in a period? Will our models continue to be accurate? This opens a lot of potential research challenges – at the end of the day we need to know how to account for such changes or at least be able to communicate how it contributes to inaccuracy in our analysis.

3.3. Terrorism Trends Over Time

Terrorism trends have changed over time. Starting in the late 1960's its apparent that terrorism has been on the rise. Particularly there was a spike in terrorism attacks between 2011 and 2015. Further analysis is warranted to see what can be attributed to a steep incline in attack incidents. We can see that between the years 2011 and 2015 (***see figure 5 below***), the preponderance of attack fatalities is carried out through Armed Assault, Assassination, Bombing/Explosions, and Hostage Taking (Kidnapping). The top category for fatalities by target type is Private Citizens & Property, Police, Military, Businesses and Government (***see figure 6 below***). It is no surprise that the top category for suicide attacks are Bombing / Explosions.

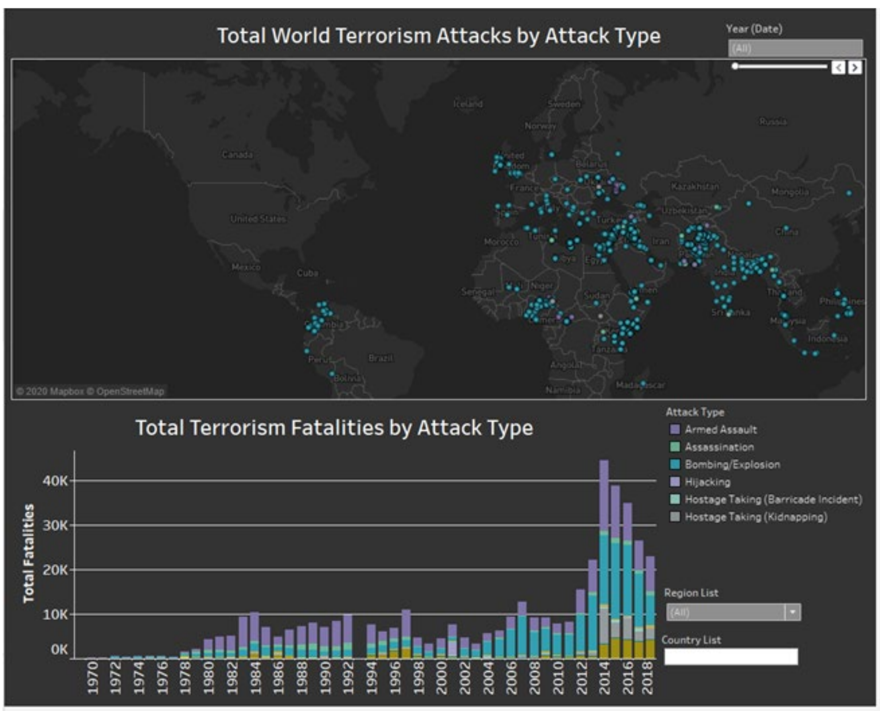

Figure 5: Terrorism Fatalities by Attack Type

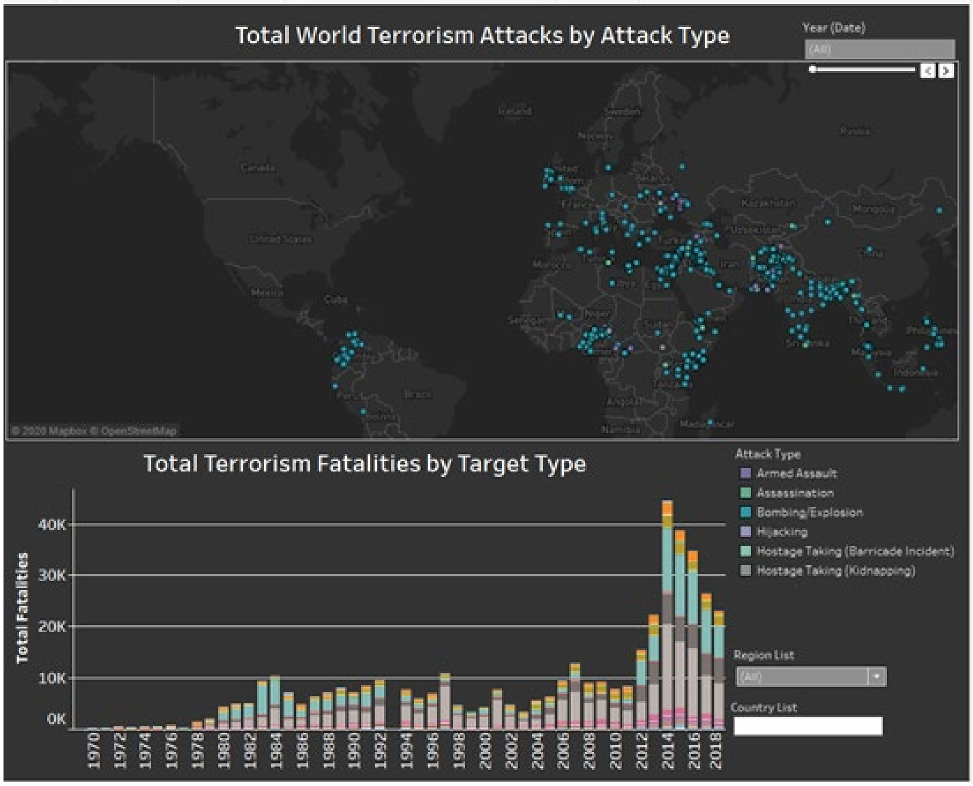

Figure 6: Terrorism Fatalities by Target Type

3.4. Geographic Trends Over Time

The Animated Total Terrorism Attacks – Heatmap (Density Visualization) which is akin to a live weather map that shows the intensity (density) of the weather i.e. more or less rain based on colors. You can see the same thing with regards to the Animated Heatmap. Looking at the animated map you can see the hotspots around the world include New York, Colombia, Argentina, Ireland, Spain, France, Algeria, Russia, Israel, Pakistan, and India (***see figure 7 below***). It is amazing to be able to see these trends within the visualizations immediately.

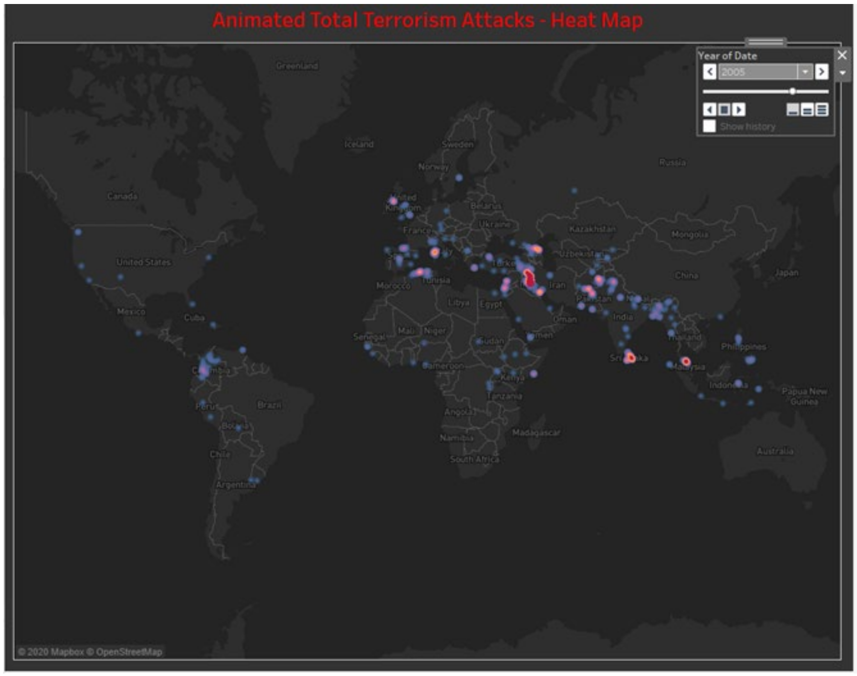

Figure 7: Animated Total Terrorism Attacks – Heatmap (Density/Intensity)

4. Implementation

The implementation of this project is currently part of a custom website developed to communicate the capstone project (***see figure 8 below***). The website is based on an HTML template utilizing CSS, JS, and other libraries. Many of the visualizations were produced in Tableau® Desktop Professional Edition (Educational License). You can really create a very professional looking website between Tableau® visualizations and custom HTML, CSS, and JS.

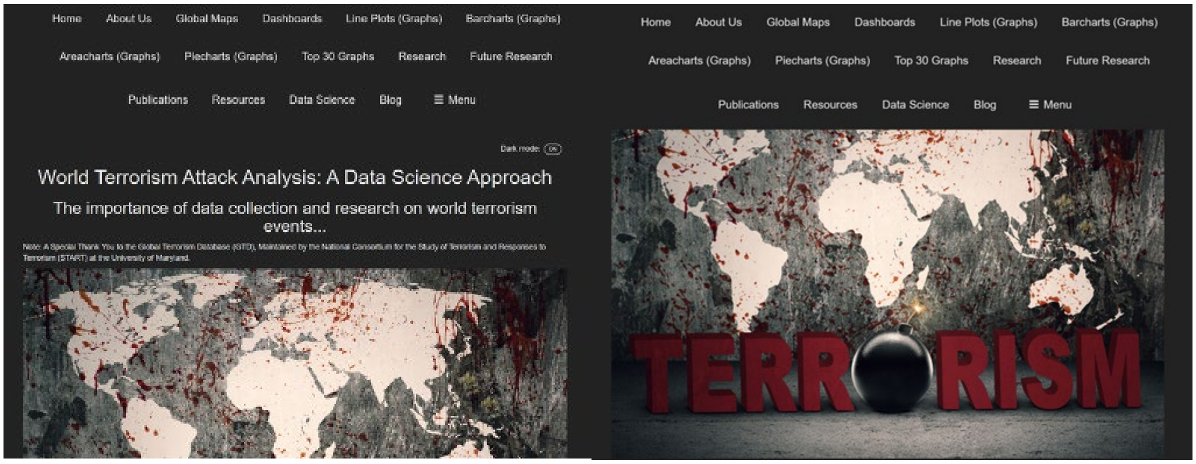

Figure 8: Homepage of the "World Terrorism Attack Analysis" Website

The classical ML models did not disappoint. I was worried that my accuracy scores below were going to be that high. I intuitively handpicked the 18 features I thought were most relevant. I did the same for the 60 features below. The main difference between the two feature lists is that for the 18 features I picked those features that were primary not secondary, so no subcategories were represented in the 18 features dataset. However, in the 60 features dataset - it included multiple subcategories and additional data elements that I thought could be important but were not as important as the primary categories. There is an exceedingly small gain between the 18 features and 60 features dataset (difference of less than 2%-4% overall) depending on the model selected. Ultimately, I picked the CART model since it was amongst the most efficient (time) and accurate. However, if I were willing to accept a slower run time for even more accuracy than the RF model with 18 features would be best. The runtimes for the 18 feature models were usually less than 5 minutes per run and the 60 feature models varied between 35 and 60 minutes. The 60-feature model was too computationally expensive and not useful.

	Features	
Models	**18 Features**	**60 Features**
CART	81.9738%	83.4496%
KNN	79.3318%	75.1204%
RF	84.5130%	84.1768%

Figure 9: F1-Score of Models

5. *Future Work*

Future work will involve adapting my research to an agnostic cloud-native environment, capable of being run on multiple platforms to maximize the return on investment through the reduction of hardware deployment time. Alternative data visualization tools and techniques will be researched for future work. How do we deal with and visualize date with regards to country borders changing or due to world events countries no longer existing? How do we deal with Terrorism Groups that fracture and morph over time? How do we track and synchronize these ontologies? Feature engineering and additional data element analysis for possible inclusion into the GTD or other databases to achieve higher fidelity predictions in various research areas. Fruitful future directions may include but are not limited to economic methodologies and research on the impact, decisions, and possible causes of terroristic activity. An interesting application of ML / DL is for the imputation of missing or incomplete data. More research needs to be done with regards to ML / DL imputation of data. The application of data science methodologies to geospatial and temporal pattern evaluation, prediction models, as well as leveraging other advanced technology and research methodologies.

6. Conclusion

This project has demonstrated that you can in fact use a model to predict a terrorist group(s) potential involvement in a terrorism incident with between approximately 82%-84.5% accuracy with a classical ML model. Visualizing this data has allowed us to uncover and compare trends that we ordinarily would not have seen. This project has also raised some dynamic challenges around country borders being reexamined, disputed or countries entirely disappearing. Terrorism Groups that fracture and morph over time, how do we account for these changes throughout our analysis? This presents some unique data challenges and innovative thinking in how to incorporate this data into further analysis and to be used in the historical context.

Another interesting challenge presented is that of the fact that current technologies do not consider historical countries that no longer exist. It is a recommendation that in the future these organizations look at incorporating historical countries into their various visualization mapping libraries not only for geo-plotting capabilities but also other libraries where appropriate for analysis. Additionally, this project has challenged us to think about ML and DL imputation methods, feature engineering and relooking our data collection methodologies. Lastly, a tool like this would enable both public and private sector organizations to predict and mitigate the risk of future terrorism attack incidents. A technology like this would surely disrupt the entire industry. A more comprehensive research approach and methodology to mature algorithms and technologies will be required – however small, this project has already demonstrated the art of the possible.

Bibliography

Abbas, Thamer, et al. “Importance and Significance of Information Sharing in Terrorism Field.” *ResearchGate*, ResearchGate, Oct. 2014, www.researchgate.net/publication/270763179_IMPORTANCE_AND_SIGNIFICANCE_OF_INFORMATION_SHARING_IN_TERRORISM_FIELD.

Agarwal, Deepak K., and Bee-Chung Chen. *Statistical Methods for Recommender Systems*. Cambridge University Press, 2016.

Alhamdani, Rafah Shihab, et al. “Recommender System for Global Terrorist Database Based on Deep Learning.” *International Journal of Machine Learning and Computing*, vol. 8, no. 6, Dec. 2018.

Bang, James Thomas, and Atin Basuchoudhary. “Predicting Terrorism: A Machine Learning Approach.” Nov. 2017.

Bowie, Neil G. “A New Inventory of 30 Terrorism Databases and Data Sets.” *Perspectives on Terrorism*, vol. 14, no. 1, Feb. 2020.

“Codebook: Inclusion Criteria and Variables.” *Global Terrorism Database*, University of Maryland, Oct. 2019, www.start.umd.edu/gtd/downloads/Codebook.pdf.

DeAngelis, Tori. “Understanding Terrorism.” *Monitor on Psychology*, American Psychological Association, 1 Nov. 2009, www.apa.org/monitor/2009/11/terrorism.

English, R. “The Future Study of Terrorism.” *Queen's University Belfast - Research Portal*, European Journal of International Security, 2016, pureadmin.qub.ac.uk/ws/portalfiles/portal/102713861/English_EJIS_Final.pdf.

“Global Terrorism Database (GTD).” *GTD*, University of Maryland, www.start.umd.edu/gtd/, (alternative citation).

“Home.” *START.umd.edu* |, University of Maryland, 11 May 2020, www.start.umd.edu/.

Jackson, Richard. “Why We Need Critical Terrorism Studies.” *E-International Relations*, E-International Relations, 8 Apr. 2008, www.e-ir.info/2008/04/08/why-we-need-critical-terrorism-studies/.

Jannach, Dietmar, et al. *Recommender Systems: An Introduction*. Cambridge University Press, 2012.

National Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland. (2019). The Global Terrorism Database™, (primary citation).

Peng, Andi. “An Integrated Machine Learning Approach to Studying Terrorism.” *Yale University*, Yale University, 2018.

Piper, Ian, et al. “Application of Microsimulation to the Modelling of Terrorist Attacks.” *University of Wollongong, Australia - Faculty of Informatics - Papers (Archive)*, 2009.

Sandler, Todd. “The Analytical Study of Terrorism: Taking Stock.” *Sage Publications*, Journal of Peace Research, 2014, journals.sagepub.com/doi/pdf/10.1177/0022343313491277.

Schuurman, Bart. “Topics in Terrorism Research: Reviewing Trends and Gaps, 2007-2016.” *Critical Studies on Terrorism*, 2019.

Singer, Gonen, and Maya Golan. “Identification of Subgroups of Terror Attacks with Shared Characteristics for the Purpose of Preventing Mass-Casualty Attacks: A Data-Mining Approach.” *Crime Science*, Jan. 2019.

Tawadros, Amira S.N. “Mapping Terrorist Groups Using Network Analysis: Egypt Case Study.” *Cairo University*, 25 Dec. 2019.

Verma, Chaman. “Predictive Modeling of Terrorist Attacks USing Machine Learning.” *International Journal of Pure and Applied Mathematics*, June 2018.