Data Analysis on the GDELT Project

Global Database of Events, Language, and Tone

Curry Buscher
University of Colorado, Boulder
Curry.Buscher@Colorado.edu

Rohan Baishya
University of Colorado, Boulder
Rohan.Baishya@Colorado.edu

Benjamin King
University of Colorado, Boulder
Benjamin.King@Colorado.edu

Gayathri Gude
University of Colorado, Boulder
Gayathri.Gude@Colorado.edu

1. PROBLEM STATEMENT/PURPOSE

The data being obtained from the GDELT project, archives a span of 45 years of "emotional snapshots". This data is essentially known to be a study of the global human society. The attributes within the dataset include information regarding societal influencers (people), risk assessment & global trends, policy reactions, and humanitarian & crisis responses. The purpose to analyzing this data is to spot trends and recognize patterns in common global events, specific to locations and human interactions. It will be investigated what events could be predicted given prior event patterns along with uncovering unexpected correlations within the data.

1. LITERATURE SURVEY

The GDELT database has been used in several prior studies. One of the ways the GDELT foundation has studied this data is to generate maps of influence in relation to 'influencers' around the world. This map connects each influencer to another based on their geographic region. This study allows researchers to determine the flow of information from a primary source throughout a region, and which influencers' input is essential to disseminate information worldwide [1]. One of the ways this was applied was to study Edward Snowden and the global effects of the continual efforts of the Wikileaks organization [2]. Next, the founder of GDELT has used this database to study wildlife trafficking in relation to global events. This study won the 2016 Wildlife Crime Tech Challenge.[3] GDELT's 'Television Explorer' system has been used by a variety of reporting organizations to analyze how the media responds to certain events. One such use analyzed Fox news' coverage of mass shootings after Robert Mueller indicted Russians in the US.[4] Another study analyzed the trend of discussing gun violence and gun control in the media, and found that after mass shootings, these conversations

spiked, then returned to a normalized level.[5] The GDELT database has also been used to track interest in various topics over time, including cryptocurrency and blockchain technologies. This summary provides an in-depth overview of these emerging technologies, and how they have come to be popularized among a general audience.[6]

2. PROPOSED WORK

To begin analyzing the data, a lot of data preprocessing will need to be done due to the sheer size and variety of data included in the 6.4 GB of GDELT data. To do this, the first step will be to pinpoint and narrow down the data to the attributes that correlations will be made for, and make sure to replace holes with projections, as well as make sure each value per attribute is homogenous based on variable type. The data will be sorted as well as deemed appropriate and specific sizes of data will need to be determined, where a proper span of information is being represented with respect to the comparisons and correlations.

The focus will be to connect historical geographic data to various attributes to make conclusions and predictions on what can happen, where it can happen, and the likelihood of these events happening at those times and places. Furthermore, we will analyze how policy changes over time have affected the various aspects of these events, as well as how different regimes and political circumstances in different countries affect social distress and protests in those areas over time.

3. DATA SET

Our dataset spans approximately 50 years of global data, and is sized at more than 6.4 GB. This is a summary of the original dataset which goes back 215 years and has an average size of 2.5 TB per year. The dataset includes numerous different attributes: **Latitude**Quantitative/Discrete, **Longitude**-Quantitative/Discrete, **Country Code**-Qualitative/Nominal, **Country Abbreviation**-Qualitative/Nominal, **Number of Events**-Quantitative/Discrete, and **Event Scale/Magnitude**-Quantitative/Continuum. The dataset

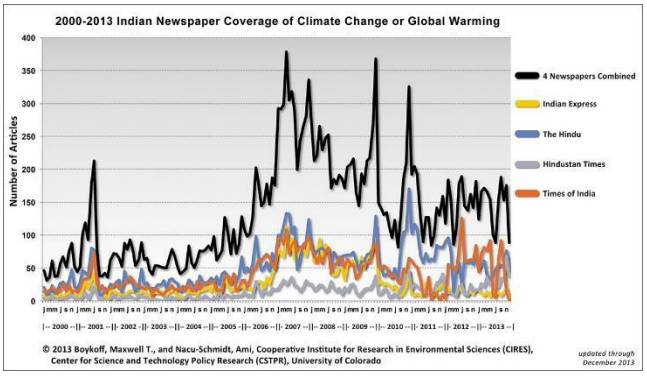
1

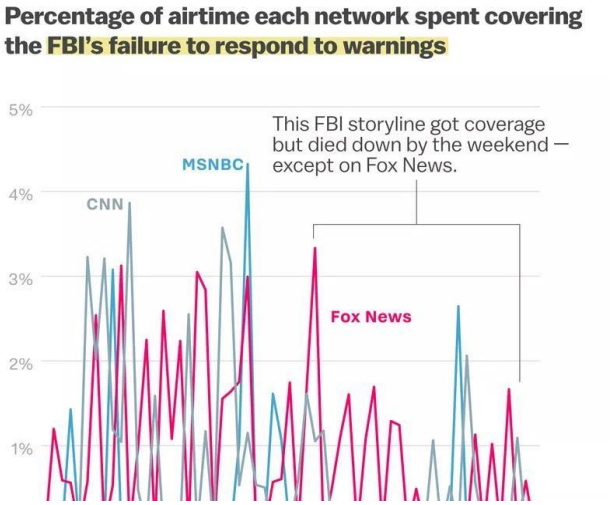
can be found on the GDELT website, and is currently stored on Ben King's computer.[7]

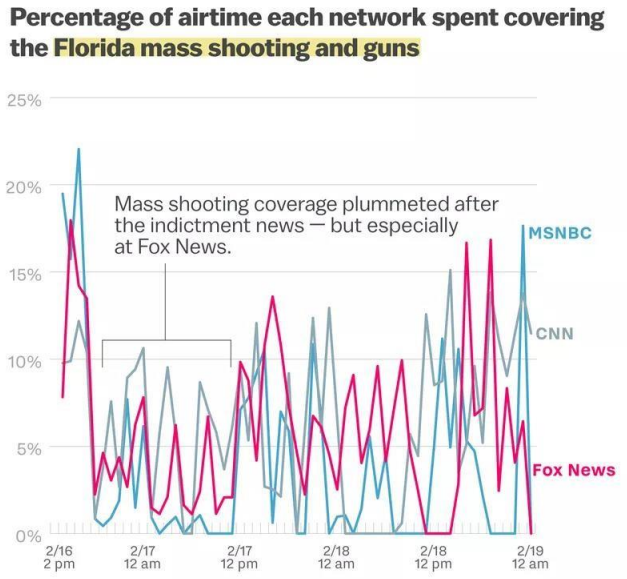
This dataset is a csv formatted package, containing all relevant datasets. This subset will then be cleansed and narrowed to specific topics, times, and attributes. From here, specific files will be mined to obtain the ultimate correlations and conclusions.

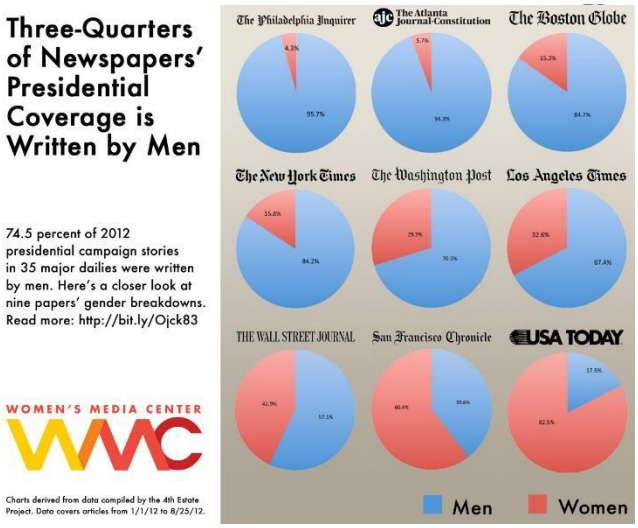
4. EVALUATION METHODS

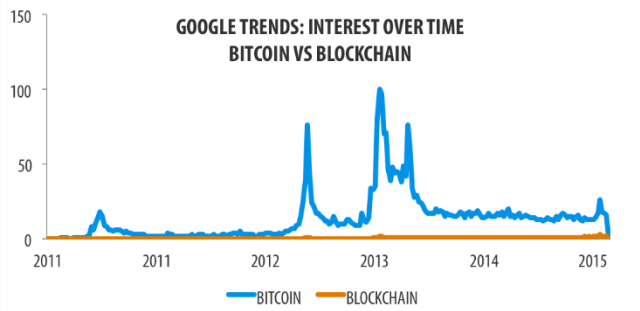
The goal of the data analysis is to map the correlations found year by year as well as group by time periods of significance. The data that will be collected and analyzed will further be compared to existing data analytics and see how they match up. Below are some data summaries in the form of visuals that comparisons will be made to and conclusions drawn with respect to our processed data:











In order to evaluate the above data and form conclusions, trends and data clustering and classification will be examined and data represented with plenty of visuals implementing MatLab plots, numpy, and pandas, after which correlations of incidents between locations and dates will be found. The analysis done will hopefully result in applicable uses such as analyzing risk assessments and crisis response times and methods to pinpoint where high risk zones are and what zones need more help and faster help. The effectiveness of policy changes will be tested as well, and predictions will be able to be made based on countless factors e.g. industry, geographical location, government, economic conditions, time, etc.

5. TOOLS

Tools that will be used include Python 3 (with a lot from the Pandas and Numpy Libraries) for preprocessing, processing, sorting, and analyzing the data throughout our project. VIsualization of data in the form of graphs and charts will be made primarily with Matlab Plots.

Tensorflow will be implemented for numerical computations and graphing our data with mathematical functions and operations as well as representing the multidimensional data arrays that connect them. Keras will be used on top of Tensorflow to enable fast experimentation and deal with configurable modules that can be plugged together with little restriction.

Git/GitHub will be used version control and to monitor/update our progress among the group. Our project link is https://github.com/cubu0178/WeLikeBigData

Finally, WEKA will be incorporated for specific, detailed data analytics including data clustering and classification as well as visualization.

6. MILESTONES: TIMELINE PLAN

The following table shows the timeline in correlation to class deadlines for each milestone. This project will be aimed to follow these given deadlines, as shown below.

Milestone	**Due Date**
Proposal Paper	March 6, 2018
Data Cleaned	March 12, 2018
Numpy & Pandas Data Exploration	March 14, 2018
Keras & Tensorflow Analysis	March 19, 2018
Weka Analysis	March 26, 2018
Graphing Results	April 2, 2018
Progress Report	April 10, 2018
Remaining work	April 16, 2018
Final Presentation	April 23, 2018
Final Paper	May 1, 2018

7. SUMMARY OF PEER REVIEW SESSION

From the presentation peer review, the current approach of finding unique trends and building from there is an appropriate path forward. The feedback received from the course coordinator was generally positive, as it referred to the nature and breadth of the current analyses, and will serve as a starting point for all future data analysis.

8. COMPLETED MILESTONES /TODO

Milestone	Due Date	Status
Proposal Paper	March 6, 2018	Completed
Data Cleaned	March 12, 2018	Completed
Numpy & Pandas Data Exploration	March 14, 2018	Completed
Keras & Tensorflow Analysis	March 19, 2018	Completed
Weka Analysis	March 26, 2018	In Progress
Graphing Results	April 2, 2018	In Progress
Progress Report	April 10, 2018	In Progress
Remaining work	April 16, 2018	To Do
Final Presentation	April 23, 2018	To Do
Final Paper	May 1, 2018	To Do

Brief Descriptions on Completed Milestones:

Data Preprocessing--Data Cleaning:

With the current dataset we have access to, we have parsed the sets we will be using for our information gathering. One of the first things to do was break the dataset into smaller, more manageable chunks. A single file of 6.5GB was too large for most programs to handle, and the computation length necessitated long time away from the computer. Furthermore, because the dataset was only locally hosted on a single computer, we were unable to perform computations on the data in multiple locations at the same time. If the program encountered a syntax error, or the computer crashed before the computation was complete, the program would have to be restarted.

Therefore, we began by splitting the dataset into individual text files of 1,000,000 entries each. This was accomplished by porting the data into a linux environment, then performing a simple *split -l 1000000* command using the unix bash. This gave us 102 files of 1,000,000 entries each, averaging ~75MB each. The leftover 103rd file is approximately 22MB, giving us a total number of entries of approximately 102,500,000. Each file was labeled according to the years spanned within (eg. 92-94, 02-03).

One interesting observation from performing this cleaning was that the density of our data was nowhere near consistent over time. For example, the first ten years of the data set (1979-1989) are contained within 5 files, meaning that each year averaged only half a million entries. In comparison, the last full year of the dataset, 2013, requires 11 separate files to be contained, an increase in size of 2200%. This disparity is largely due to the increase in the number of reporting countries (sources) over time, as well as the number of captured events per day.

For comparison, on January 1st, 1979, the database captured 306 event summaries from 91 different sources. On January 1st, 2013, the database captured 15,063 event summaries from 1,003 different sources.

Generally, the quality of data is consistent over time. For comparison, the first file containing the years 1979-1982 is missing 67,408 values for SourceGeoType, while the first file of 2013 is missing 62,193. However, the data is much more thorough as one moves forward in time, and the reporting from each source is improved. On January 1st, 1979, the highest number of events came from the US government at 28 separate events. On January 1st 2013, 91 different countries reported more than 28 events.

Data Preprocessing—Data Transformation:

One of the most difficult things about working with the GDELT database is the extensive list of codes which must be matched in order to fully understand any given event. The CAMEO Code dictionary must be used in order to decipher each code. The current GDELT database uses the CAMEO Version 1.1b3 Manual, which is almost 200 pages long, and can be found online in a pdf. This dictionary lists various codes which can be references to people, places, governments, religions, and organizations which each share certain codes.

For example, the Cameo Code [VATGOV] would refer to either Pope John Paul II, Pope Benedict XVI, Pope Francis I, the current pope (generally), or Vatican Officials. Because there are millions of events and each one never fits neatly into a single category, it is necessary to keep these codes and what they represent separate. However, when examining data and trends, it is often necessary to convert these codes into what they represent in order to determine what specifically each event is referring to.

Another example which shows the difficulties of deciphering these codes is the code [USA], which is the most frequently-appearing code in the dataset. This code can refer to any of the following according to the CAMEO dictionary: the United States, a US-led coalition, the University of South Florida, US Military bases in Cuba (generally), the state of New York, the state of New Mexico, the city of Los Angeles, Guantanamo Bay, or an American Bomb Disposal Expert.

Data Preprocessing—Data Reduction:

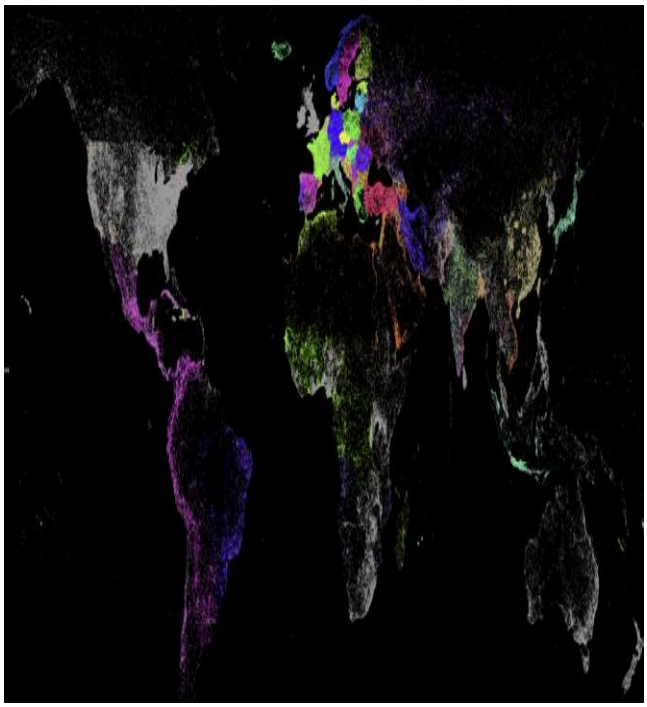
Because we are already using the 'reduced' GDELT dataset, our data has already been reduced to its simplest form. For example, the full 2015 GDELT dataset is comprised of more than 2.5TB of data. This can be compared to the 35 years of data present in our dataset, which has been reduced to a mere 6.5GB. Therefore, we feel as if our data has been reduced to its simplest form while also retaining meaningful properties.

9. RESULTS SO FAR AND GOALS:

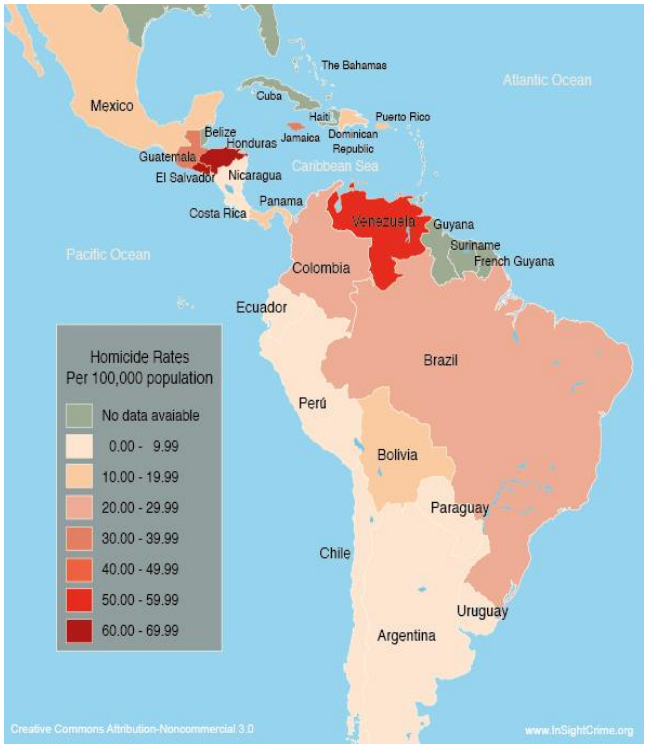
All results and data compiled to this point in time, has been purely unsophisticated. However, we are hoping to compile our results throughout all dataset attributes to get a more meaningful overlook into the underlying, hidden meanings within the data. We hope to find strong correlations between worldwide events, location and incident specific.

Some graphs we hope to develop through the mining of our data, include worldwide correlations (depicted through color related graph) and incident number events (depicted through frequency). Some examples include: (1) A graph of a worldwide scan of similar incidents and (2) A graph of frequency of a certain event in a certain region.

(1)



(2)



10. REFERENCES:

[1] The GDELT Project. (n.d.). Retrieved March 06, 2018, from https://www.gdeltproject.org/solutions.html#influencers

[2] Leetaru, K. (2013, December 31). King Snowden and the Fall of Wikileaks. Retrieved March 06, 2018, from http://foreignpolicy.com/2013/12/31/king-snowden-and-the-fallof-wikileaks/

[3] Wildlife Trade News - Winners of Wildlife Crime Tech Challenge announced. (n.d.). Retrieved March 06, 2018, from http://www.traffic.org/home/2016/20/winners-of-wildlife-crimetech-challenge-announced.html

[4] Chang, A. (2018, February 19). Fox Newss appalling past 72 hours, analyzed. Retrieved March 06, 2018, from https://www.vox.com/2018/2/19/17027456/fox-news-muellerindictment-trump

[5] Qureshi, H. (2018, February 21). Conversations about gun violence spike when shootings occur, then decline soon after.
Retrieved March 06, 2018, from http://www.dailyhelmsman.com/news/conversations-about-gunviolence-spike-when-shootings-occur-thendecline/article_d938d98c-1754-11e8-9c61-4b1b27430f24.html [6] Tracking Bitcoin/Cryptocurrencies

And Blockchain Using GDELT Summary. (2018, January 08). Retrieved March 06, 2018, from https://blog.gdeltproject.org/tracking-bitcoincryptocurrencies-blockchain-using-gdelt-summary/ [7] The GDELT Project. (n.d.). Retrieved March 06, 2018, from https://www.gdeltproject.org/data.html#rawdatafiles
[8] Who's Writing Nine Newspapers' Presidential Election Coverage? (n.d.). Retrieved March 06, 2018, from http://vitaminw.co/politics/whos-writing-nine-newspaperspresidential-election-coverage
[9] Inside the Greenhouse. (n.d.). Retrieved March 06, 2018, from http://sciencepolicy.colorado.edu/icecaps/research/media_coverag e/india/index.html