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. ABSTRACT

The GDELT database is a global event database spanning more than 30 years, containing entries surrounding politics, economics, war, and peace. Our project revolved around the reduced event database, which is a smaller, summarized version of the original. This version contains the same entries as the complete database, while combining each entry of a similar type occurring on the same day into a single entry. This allows for much easier analysis of the data without compromising on the underlying trends and patterns.

2. INTRODUCTION

We strove to analyze and examine, in detail, the various event trends and reactionary data of human society. This data is being obtained from the GDELT project, in which our data 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 our dataset, include information regarding societal influencers (people), risk assessment & global trends, policy reactions, and humanitarian & crisis responses.

We sought to answer questions surrounding global cooperation and conflict over time as they relate to events in the dataset, including:

a) Was the breakup of the Soviet Union preceded by increased cooperation or conflict with other nations? How did the breakup affect US-Russian relations and relations with other nations?

b) Do the Olympics have any effect on global cooperation? Specifically, does the host nation of the Olympics have more cooperation than other nations?

c) How have the world's most threatening nations evolved over time?

3. 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.

4. RELATED WORK

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]

5. PROPOSED WORK (PRIOR TO PROJECT 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.

6. DATA SET

Our dataset spans approximately 35 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 is in numerous locations: Our local computers, Google Drive, GitHub, and can be found at the GDELT Website: https://www.gdeltproject.org/data.html.

The reduced event dataset compiles each event reported in a location per day, and conglomerates them into a single entry. For example, each event of type 'Protest' in location 'Russia' is tallied and incorporated into one entry.

A list and summary of the attributes are as follows:

A) **Date -** YYYYMMDD formatted date of each event

B) **Source -** Source of event, using appropriate actor code eg. [RUS]. eg. Russia bombs a city in Afghanistan. The source is Russia.

C) **Target -** Target of event, using appropriate actor code eg. [IRN]. eg. The United States bombs a city in Iraq. The target is Iran.

D) **CameoCode -** Refers to the type of event being recorded. This is pulled from a list of 300+ distinct event types ranging from:
01: A public statement being made
To
203: Engaging in Ethnic Cleansing

This code represents a distinct event type and determines which events are combined together.

E) **NumEvents -** Lists the total number of events of a given type occurring in a country on a given day eg. 12 protests in India. NumEvents is 12.

F) **NumArts -** Lists the total number of articles referenced about a given type of event. Eg. There are 45 different articles written on a given day about the chemical weapons used in Syria. NumArts is 45. This attribute correlates roughly the confidence with which the GDELT project is confident about an event occurring. If an event has a very low NumArts value, it is much more likely to be a smaller event in scale, while a very high NumArts value implies the opposite.

G) **QuadClass -** This attribute classifies each CameoCode into one of four categories:
1 - Verbal Cooperation eg. Verbal Promises / Agreements
2 - Material Cooperation eg. Sending Aid
3 - Verbal Conflict eg. Threats
4 - Material Conflict eg. Military Action

This attribute is one of the most useful in our dataset, as it allows the collapse of the hundreds of CameoCodes, of which there might be only 2 or 3 examples of each per year, into 4 categories, of which there are millions of examples. This allows for more rigorous analysis over the entirety of our data.

H) **GoldStein -** This attribute assesses the relative impact of an event based on its event type. The Goldstein value is a number from -10 to +10, which scores the type of event based on its effect on the stability of a country. A negative Goldstein score implies that the type of event is destabilizing, while a positive Goldstein score implies that the event tends to stabilize the location in which it occurs. This attribute only ranks each CameoCode, and does not rank or compare individual events. For example, a riot of 100 people and a riot of 10,000 people will have the same Goldstein score, because both are in EventType ‘Riots’.

I) **SourceGeoType -** Resolution of the Source Lat/Long data. For example, a resolution of 1 would correspond to a lat/long of only whole numbers. A resolution of 4 would correspond to a lat/long of a whole number, followed by 4 decimals.

J) **SourceGeoLat -** Source Latitude of the first actor.

K) **SourceGeoLong -** Source Longitude of the first actor.

L) **TargetGeoType -** Resolution of the Target Lat/Long data.

M) **TargetGeoLat -** Target latitude of the second actor.

N) **TargetGeoLong -** Target longitude of the second actor.

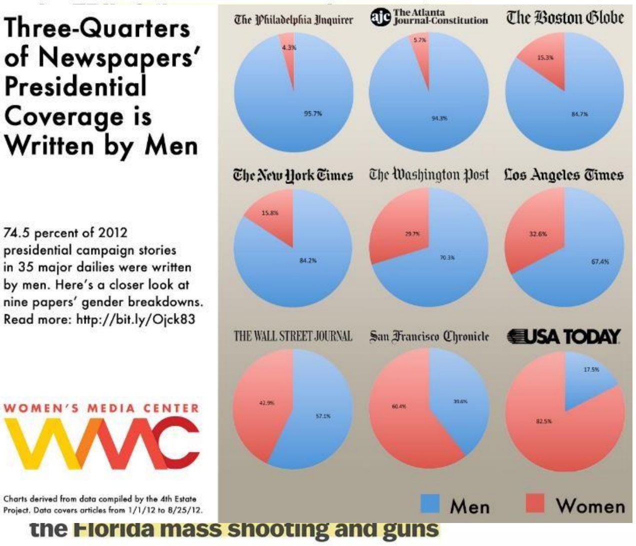
O) **ActionGeoType -** Resolution of the Action Lat/Long data

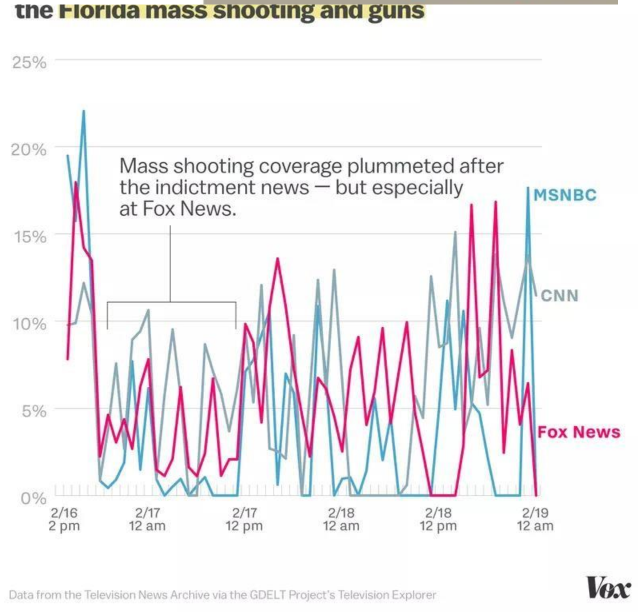
P) **ActionGeoLat -** Latitude of the location of the action/event.

Q) **ActionGeoLong -** Longitude of the location of the action/event.

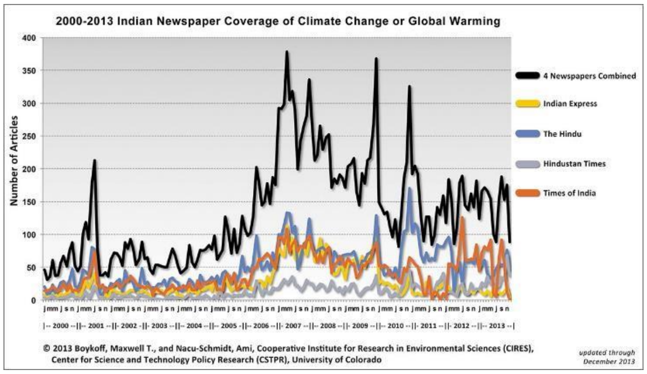
7. EVALUATION METHODS (PRIOR TO PROJECT WORK):

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.

8. TOOLS

Tools that will be used include Python 3 (utilizing 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.

One of the primary tools utilized was the Dask library within an ipython notebook framework. Dask is a library which allows for the conversion of very large datasets into a series of pandas dataframes, which are each an individual component of a Dask dataframe. The Dask library API contains a subset of all commands found in the pandas data frame library, and performs functions in a similar manner. This library is useful in that it allows one to sequentially import data into data frames rather than trying to store the entire dataset in memory at once. However, Dask has its downsides.

For example, Dask explicitly bans some of the more expensive functions found in the pandas library, such as replace and drop, which make data cleaning very difficult. Furthermore, even though Dask utilizes much of the same codebase as pandas, because of the large dataset sizes involved, the commands take much longer to run.

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

9. MAIN TECHNIQUES APPLIED

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.

10. MILESTONES: TIMELINE PLAN

The following table shows the timeline in correlation to class deadlines for each milestone. This project will was 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
Graphing Results	April 2, 2018
Progress Report	April 10, 2018
Remaining work	April 16, 2018

Final Paper	April 30, 2018
Final Presentation	May 2, 2018

11. SUMMARY OF PEER REVIEW SESSION

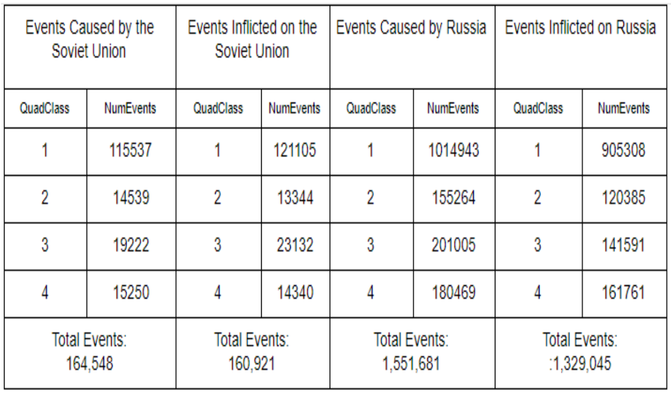
From the presentation peer review, the current approach of finding unique trends and building from there is an appropriate path forward. The feedback received from the course coordinator was generally positive, as it referred to the nature and breadth of the current analyses, and will serve as a starting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

12. KEY RESULTS AND VISUALIZATIONS

****Refer back for a refresher on QuadClass meanings on page 3 *****

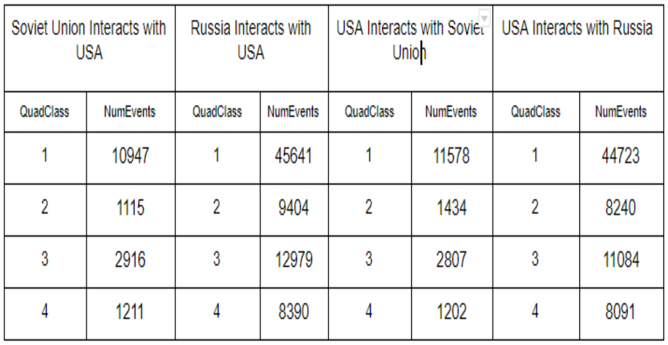
(1) The first question we sought to answer can be proved by the following two datasets.

The following is the data we collected when investigating cooperation and conflict events in the Soviet Union (pre-1992) versus Russia (1992 - present).

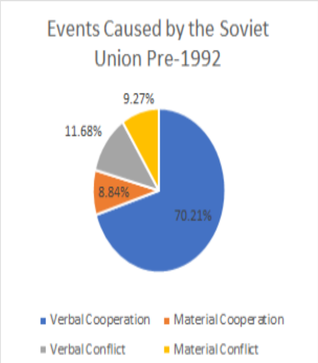


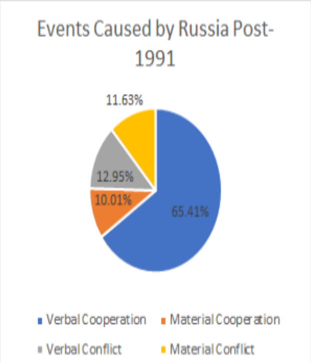
Events Caused by the Soviet Union		Events Inflicted on the Soviet Union		Events Caused by Russia		Events Inflicted on Russia	
QuadClass	NumEvents	QuadClass	NumEvents	QuadClass	NumEvents	QuadClass	NumEvents
1	115537	1	121105	1	1014943	1	905308
2	14539	2	13344	2	155264	2	120385
3	19222	3	23132	3	201005	3	141591
4	15250	4	14340	4	180469	4	161761
Total Events: 164,548		Total Events: 160,921		Total Events: 1,551,681		Total Events: :1,329,045	

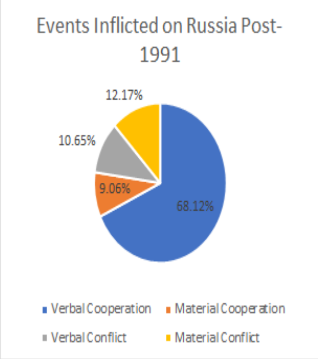
This second set of data alludes back to how the US-Russian relations were affected by the breakup.

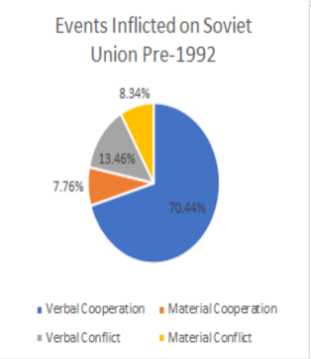


Soviet Union Interacts with USA		Russia Interacts with USA		USA Interacts with Soviet Union		USA Interacts with Russia	
QuadClass	NumEvents	QuadClass	NumEvents	QuadClass	NumEvents	QuadClass	NumEvents
1	10947	1	45641	1	11578	1	44723
2	1115	2	9404	2	1434	2	8240
3	2916	3	12979	3	2807	3	11084
4	1211	4	8390	4	1202	4	8091



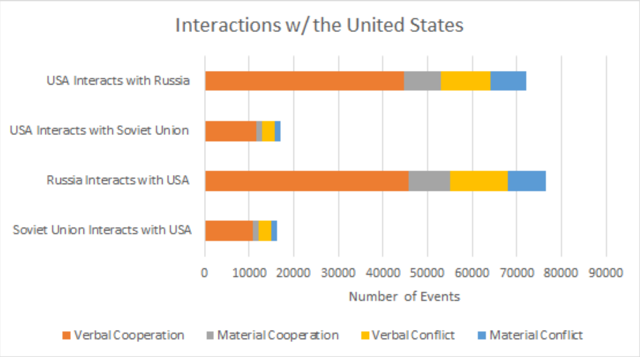






Although the number of events in each category and in total increases dramatically over time with modern day Russia relative to the Soviet Union pre-1992, the proportion of types of events remains almost the same over time. On another note a recurring trend over time is a large amount of verbal cooperation over time both delivered and received by Russia (circa 65-70% of events) but without much follow through with actual material cooperation (only 7-10%).

In contrast while verbal conflict is relatively low overall (10-13%), these verbal conflicts are largely followed through with material conflict (8-12%).

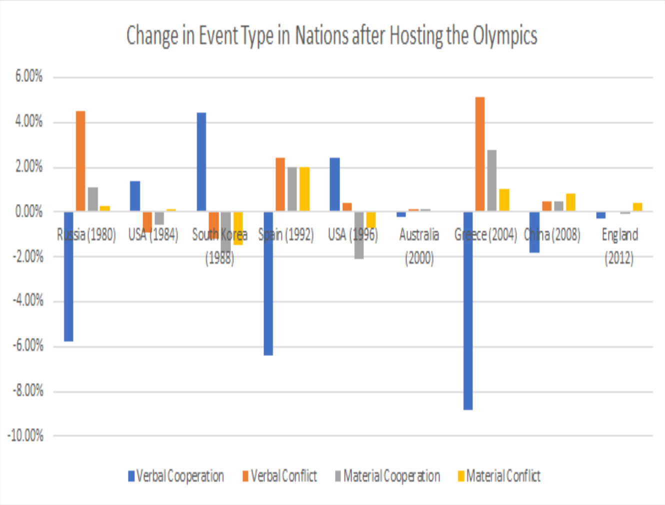


This data depicts the number of interactions on both ends go up significantly after the fall of the Soviet Union. However, the relative proportions of types of events remains relatively the same. Cooperation is mainly all verbal, with little material cooperation to back it up. Meanwhile, there is little conflict in comparison. However, the proportion of material conflict to verbal conflict is almost 1:1, which entails verbal disagreements were often followed up with malevolent actions against one another, while cooperative agreements were not followed up with cooperative actions.

In conclusion, cooperation with both the Soviet Union and Russia by the United States, both delivered and received, is all talk and no action, while verbal conflict is almost always followed up by actual material conflict. In terms of number of actual material actions, conflict and cooperation are comparatively approximately the same.

(2) The second question we sought to answer can be proved by the following dataset.

QUADCLASS	Russia (1980)	USA (1984)	South Korea (1988)	Spain (1992)	USA (1996)	Australia (2000)	Greece (2004)	China (2008)	Great Britain (2012)
1	-5.8%	1.4%	4.4%	-6.4%	2.4%	-0.2%	-8.8%	-1.8%	-0.3%
3	4.5%	-0.9%	-1.2%	2.4%	0.4%	0.1%	5.1%	0.5%	0.0%
2	1.1%	-0.6%	-1.8%	2.0%	-2.1%	0.1%	2.8%	0.5%	-0.1%
4	0.3%	0.1%	-1.5%	2.0%	-0.7%	0.0%	1.0%	0.8%	0.4%



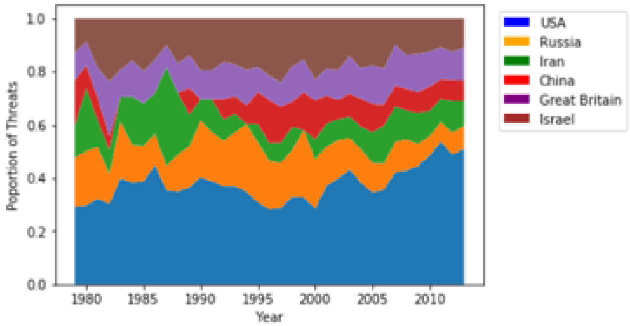
Data that will not be considered in evaluating correlations include the 2000 Olympics in Australia and the 2012 Olympics in England as there is almost no change in any sort of event type.

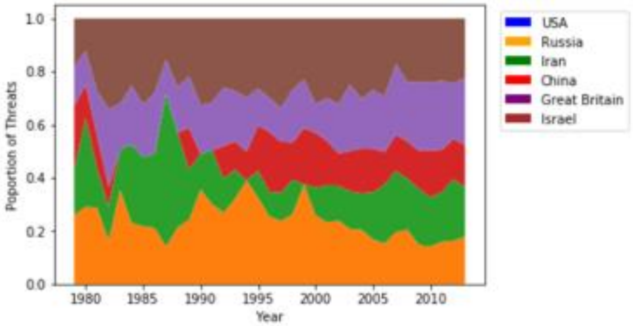
The largest changes occurred with verbal cooperation during the other Olympics; the United States and South Korea were the only nations that had increases in verbal cooperation after hosting the Olympics, while Russia, Spain, Greece, and China all showed drops in verbal cooperation. Material Cooperation was the opposite in these scenarios, with the U.S. and South Korea showing decreases in it, while the other nation showed increases. This shows an inverse correlation between verbal and material correlation, as it seems Russia, Spain, Greece, and China followed through on verbal agreements with material cooperative actions, closing the gap between verbal and material cooperation, while the U.S. and South Korea made more verbal agreements without as much cooperative material action.

Russia, Spain, and Greece had the only significant increases in verbal conflict events when they hosted the Olympics. However, Russia's and Greece's actual material conflicts did not follow, barely increasing and not nearly as much as the verbal disagreements they had. In contrast, Spain had almost as much of an increase in material conflicts as they did in verbal conflicts, making them the only one of the three countries that acted on most of their new verbal disagreements with material conflict.

South Korea was the only nation that had a notable decrease in either verbal or material conflict, with both following each other fairly closely, which may relate to a tad more peace within the nation after hosting the Olympics.

We also investigated the change in counts and proportions of threats issued by the top six issuing nations from 1979 to 2015. Few threats issued before 1995, after which the USA increased its number of threats issued, especially following 2001. In 2005, Threats issued around the world begin to increase substantially to number in the thousands per state.

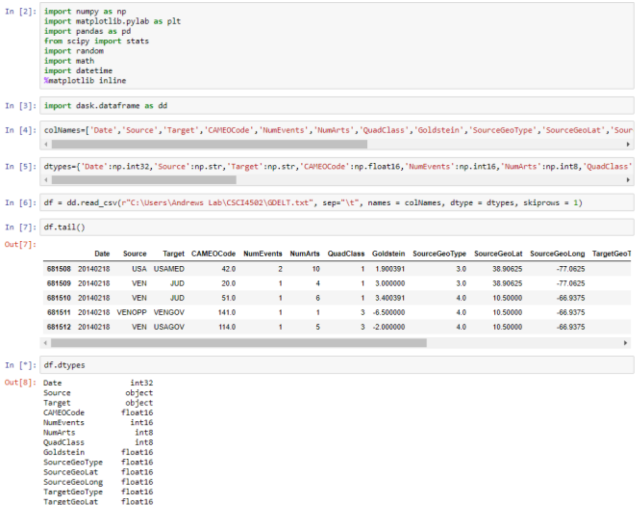




In conclusion, cooperation with both the Soviet Union and Russia by the United States, both delivered and received, is all talk and no action, while verbal conflict is almost always followed up by actual material conflict. In terms of number of actual material actions, conflict and cooperation are comparatively approximately the same.

13. BASELINE CODE SNIPPET

The following image shows parts of the GDELT dataset being read in through each specified attribute. The dataset was then furtherly formatted to analyze and mine the information/results presented in the previous section. This code was written in Python, and ran through the tools: Jupyter Notebooks and Anaconda.



```
In [2]: import numpy as np
        import matplotlib.pylab as plt
        import pandas as pd
        from scipy import stats
        import random
        import math
        import datetime
        %matplotlib inline

In [3]: import dask.dataframe as dd

In [4]: colNames=['Date','Source','Target','CAMEOCode','NumEvents','NumArts','QuadClass','Goldstein','SourceGeoType','SourceGeoLat','Sour

In [5]: dtypes={'Date':np.int32,'Source':np.str,'Target':np.str,'CAMEOCode':np.float16,'NumEvents':np.int16,'NumArts':np.int8,'QuadClass'

In [6]: df = dd.read_csv(r"C:\Users\Andrews Lab\CSCI4502\GDELT.txt", sep="\t", names = colNames, dtype = dtypes, skiprows = 1)

In [7]: df.tail()

Out[7]:
                  Date  Source  Target  CAMEOCode  NumEvents  NumArts  QuadClass  Goldstein  SourceGeoType  SourceGeoLat  SourceGeoLong  TargetGeoT
        681508  20140218     USA  USAMED       42.0          2       10          1   1.900391            3.0      38.90625       -77.0625
        681509  20140218     VEN     JUD       20.0          1        4          1   3.000000            3.0      38.90625       -77.0625
        681510  20140218     VEN     JUD       51.0          1        6          1   3.400391            4.0      10.50000       -66.9375
        681511  20140218  VENOPP  VENGOV      141.0          1        1          3  -6.500000            4.0      10.50000       -66.9375
        681512  20140218     VEN  USAGOV      114.0          1        5          3  -2.000000            4.0      10.50000       -66.9375

In [*]: df.dtypes

Out[8]: Date               int32
        Source            object
        Target            object
        CAMEOCode        float16
        NumEvents          int16
        NumArts             int8
        QuadClass           int8
        Goldstein        float16
        SourceGeoType    float16
        SourceGeoLat     float16
        SourceGeoLong    float16
        TargetGeoType    float16
        TargetGeoLat     float16
```

14. APPLICATIONS

Our analysis sought to find correlations between global events and the amount of cooperation or conflict between nations and regions. By asking these questions and analyzing the results, one might be able to predict when and where conflicts arise to avoid them, or show that increasing cooperation among nations helps to alleviate the amount of negative events that occur around the world.

Generally, this knowledge is only of value when placed in the hands of global leaders and influencers. For example, knowing that a conflict is about to occur between Russia and the United States is of little value unless one is in a position to alter the outcome of such a conflict or prevent it altogether. Furthermore, our main questions aimed to analyze the effects of communism and instability on the relationships between two nations. If one can definitively prove that Soviet Russia was beholden to more conflicts when it was a communist nation, and more positive events when it was not, this could serve as a warning to other communist nations around the world.

In the case of the break-up of the Soviet Union, one can analyze whether or not, prior to the break-up, there was an increase in conflict with other nations, or whether there was an increase in cooperation. This could serve as a guidebook for how to deal with rogue states; one can either pressure / threaten them into submission, or try to work together and find common ground. Either way, the information can be valuable when analyzing the rise and fall of similar nations today.

Overall, in regard to threats issued around the world, the number of threats issued by the top six threat issuing nations has remained about constant every year until a massive continuing increase leading into the 21st century. The USA has increased its proportion of threats from one-third to nearly a half of all threats issued by the top six nations. Knowing such information would prove to be beneficial for the nations' security.

15. 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

[10] Dask 0.17.2 Documentation, Anaconda, 2014, dask.pydata.org/en/latest/dataframe-api.html