Max Herrera, Joshua Herrera, Patrick McDevitt, and Sunna Quazi

MSDS 7330 File Organization and Database Management

Section 405, Group A

Global Terrorism Database

- Open-source database including information on terrorist events around the world
- Terrorist Attacks from 1970-2015
- 156,772 terrorist attacks
  - More than 83,000 bombings, 18,000 assassinations, and 11,000 kidnappings since 1970
- 137 Attributes
- Systematic data on domestic as well as international terrorist incidents
- Over 4,000,000 news articles and 25,000 news sources were reviewed to collect incident data from 1998 to 2016 alone
- https://www.kaggle.com/START-UMD/gtd. (29 Mb)
- https://www.start.umd.edu/gtd

Currently the most comprehensive unclassified database on terrorist events in the world

Team Project

- Identified GTD as opportunity to explore / organize / normalize / evaluate a large dataset in support of objectives of this course
- Identified relational database as most supportive for this database and project
- Convert raw data set to usable tables
- Imported to SQL database
- Normalized database
- Identified time-based and geographic insights

GTD defines Terrorism as:

"The threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation."

Feel the Data

Before loading the data, it's a good idea to get a summary of what we are looking at

Below is a text analysis performed in excel on the top 200 words in the dataset

Observable Trends

- 1970s: Catholics vs. Protestants, US, Puerto Rico
- 1980s: Abortion, Israel
- 1990s: Turkey, Israel, India

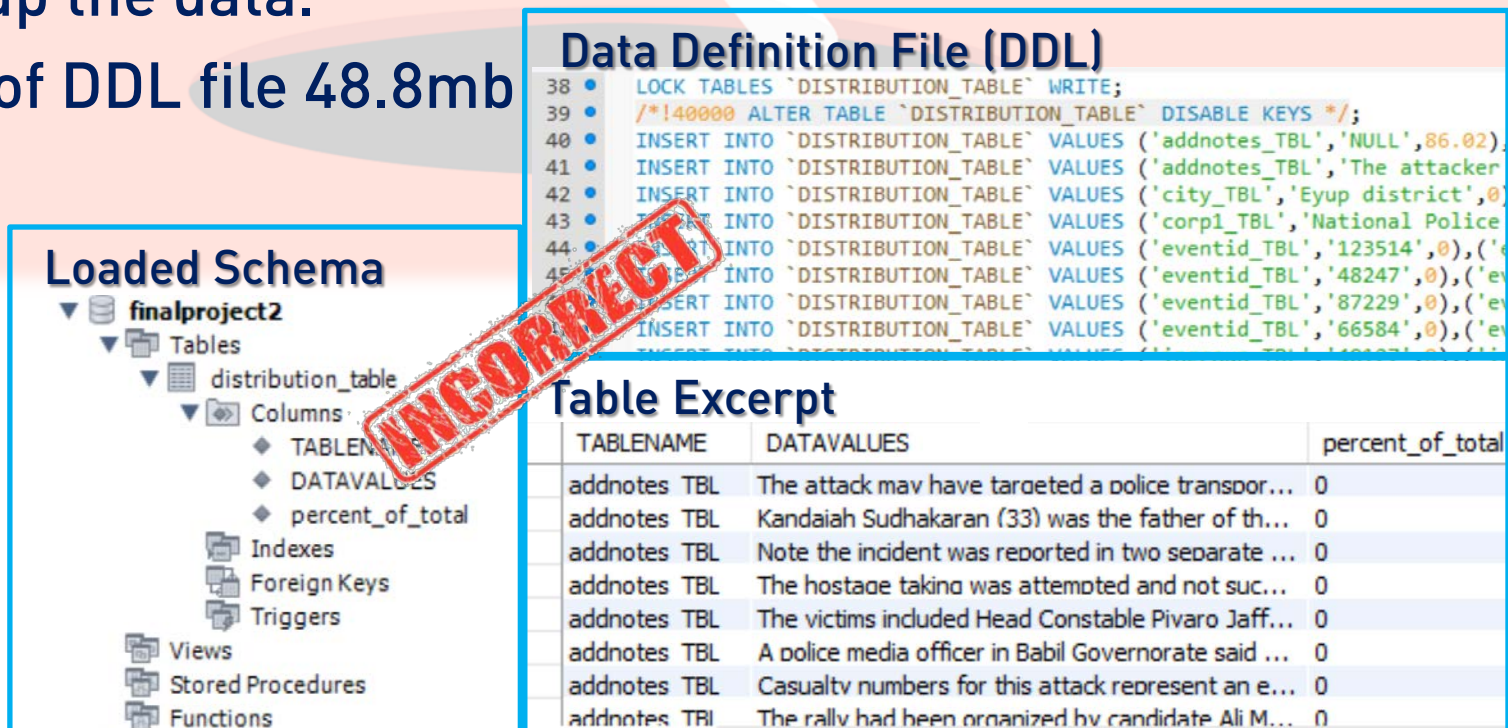
n_1970	n_1972	n_1974	n_1976	n_1978	n_1980
1970	firearm	explosive	explosive	explosive	explosive
united	explosive	firearm	incendiary	firearm	firearm
government	catholic	catholic	firearm	automatic	automatic
bomb	bomb	incendiary	automatic	incendiary	military
office	protestant	bomb	office	pistol	unit
1971	1972	protestant	protestant	office	pistol
times	york	automatic	bomb	police	incendiary
york	times	bank	pistol	station	police
6	sniper	1975	u.s	firearms	office
u.s	incendiary	office	catholic	bomb	station
august	automatic	york	pistols	bank	bomb
explosive	u.s	times	bank	guard	town
police	bank	police	vehicles	party	bank
building	john	1974	residence	patrol	bus
committee	police	america	firearms	store	car
printing	car	john	police	pistols	post
riots	james	puerto	political	1978	patrol
civil	patrol	international	1976	building	offices
senate	foot	security	vehicle	post	embassy
criminal	19	christopher	embassy	bus	store
disorders	1973	pistols	york	offices	grenade
operations	black	political	party	radio	firearms
protest	embassy	consulate	station	national	radio
perpetrators	january	firearms	consulate	civil	national
casualties	office	terrorism	offices	political	dynamite
washington	william	u.s	building	car	power
america	david	violence	puerto	home	grenades
post	political	embassy	national	chief	street
university	20	2005	car	center	building
california	bombs	4	john	government	line
bombed	firearms	chronology	store	military	rifle
bank	united	hewitt	times	residence	vehicle
4	booby	modern	19	town	city
fire	trap	praeger	hotel	puerto	guard
los	zebra	station	bldg	rifle	tower
security	8	store	william	jose	center
terrorism	pistol	bombs	owned	army	jose
april	set	building	2	embassy	party
political	22	pistol	america	leader	leader
angeles	thomas	william	frg	de	1981

Why Choose SQL?

1. We didn't know about MongoDB until last week ;)
2. Raw dataset was in tabular / column structure
3. The 137 attributes are interdependent, which complement an relationship database
4. There was seemingly a great amount of redundant data, which would benefit from normalization
5. Relational database facilitates analyzing the multiple subtypes of information: geographic, forensic, demographic, political

Data Loading Pitfalls

- Initially, the GT data would not import into MYSQL
- We needed to make each column varchar(MAX) initially. Each column was surrounded by triple quotes, so we cleaned up the data.
- File size of DDL file 48.8mb



Data Wrangling

1. Import the data into a Microsoft SQL Server database.

We needed to make each column varchar(MAX) initially. Each column was surrounded by triple quotes, so we cleaned up the data:

2. Write scripts to determine the max length of each column, and alter the column width for each of the 127 columns accordingly.
3. Use the sliding window functions to determine distribution of values for each of the 127 columns.
4. Output the distribution of values per column into one table, DISTRIBUTION_TABLE.
5. Connect the SQL server to MYSQL to transfer the data using SSIS via an ODBC connection.
6. Perform a MYSQL dump of the data for the GT (I.e., the global terrorism table) and the DISTRIBUTION_TABLE and uploaded it into box so that it could be loaded by the other teammates.

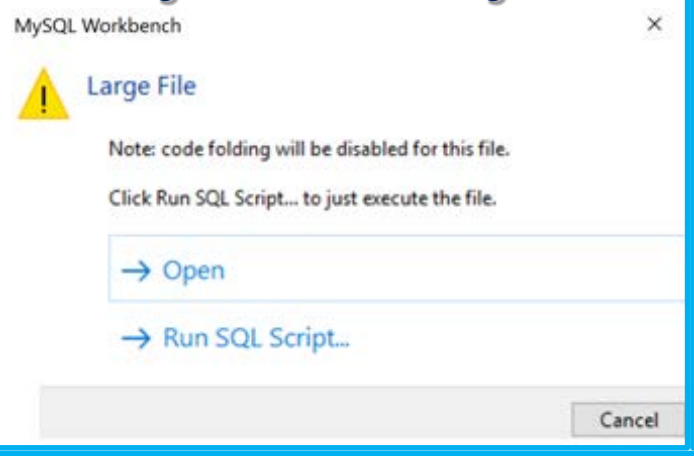
   This allowed the team to start with the same dataset.

All the scripts that were created to wrangle, and analyze the data can be found here. https://smu.box.com/s/zhbm2idikopnuhxfvcxtmc880nud5pxa. There were 22 scripts written to wrangle and profile the data. The SQL dump generated to scripts, which can be found in the importdataintomysql folder. The stored procedures and sql used is for both microsoft sql server and mysql.

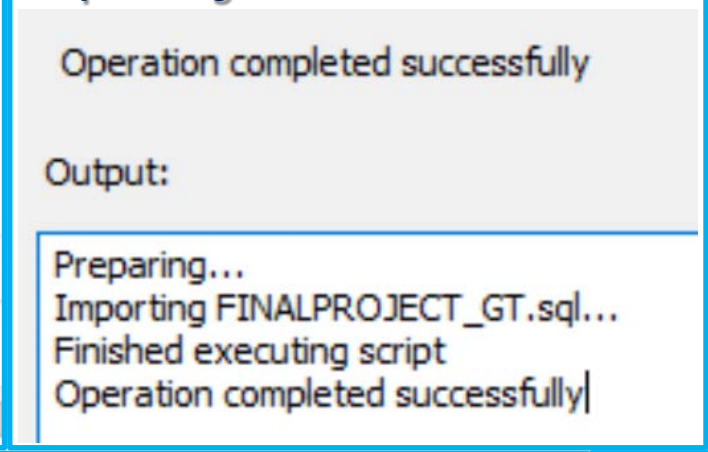
Our Framework

Final File size is 201mb (from 29mb)

Loading Data- Too Large



Importing



Loaded Schema

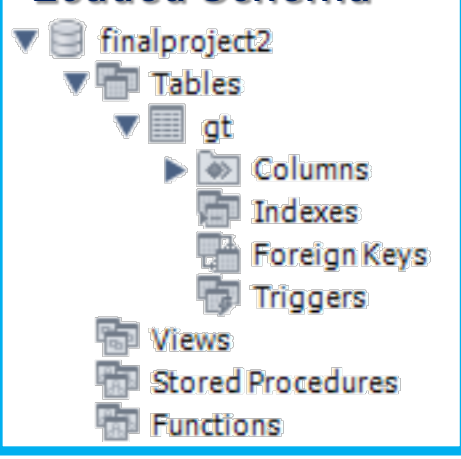
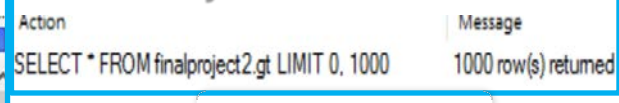


Table Excerpt

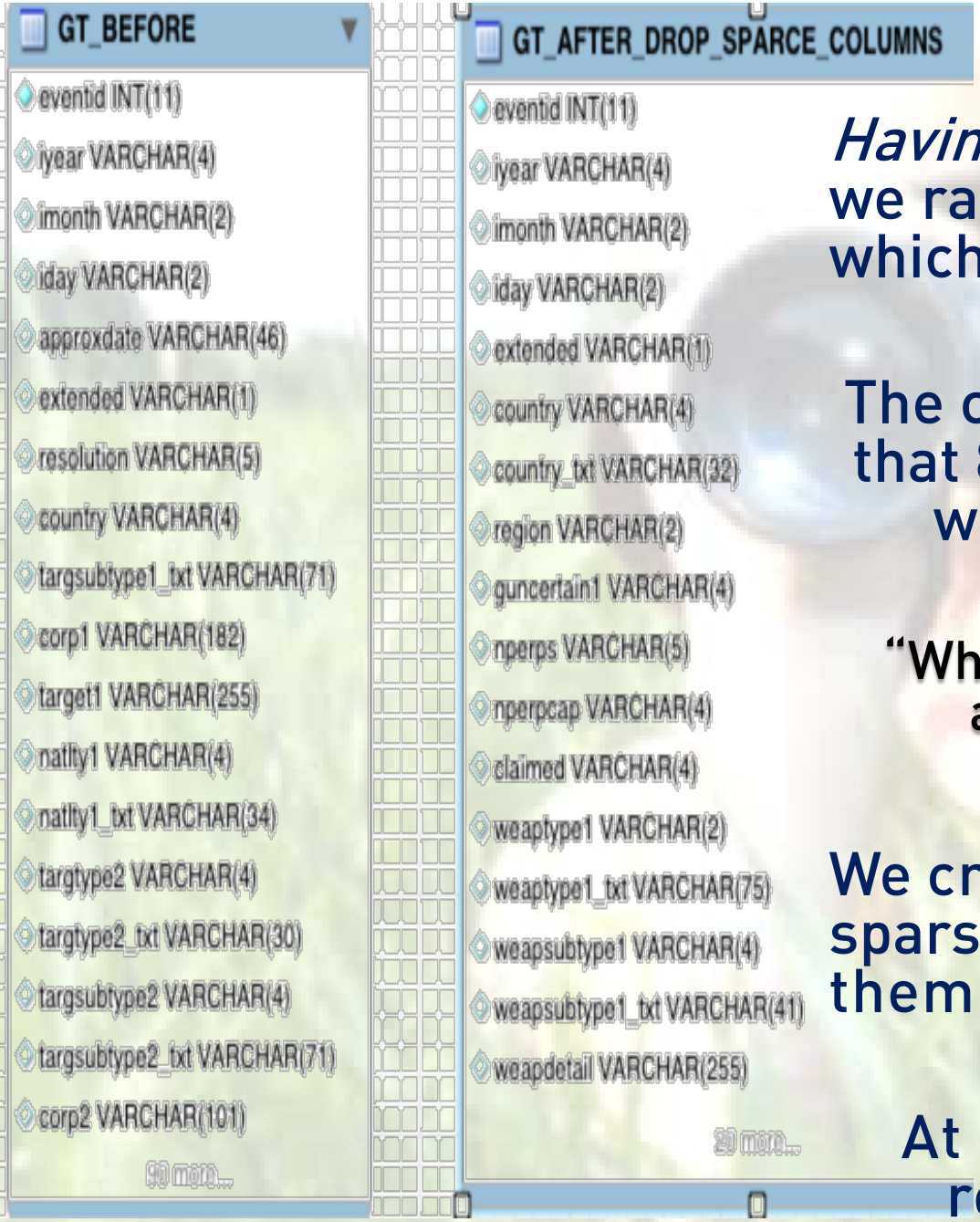
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

eventid	iyear	imonth	iday	approxdate	extended	resolution	country	country_txt	region	region_txt
1	1970	0	0	NULL	0	NULL	58	Dominican Republic	2	Central America & C
2	1970	0	0	NULL	0	NULL	130	Mexico	1	North America
3	1970	1	0	NULL	0	NULL	160	Philippines	5	Southeast Asia
4	1970	1	0	NULL	0	NULL	78	Greece	8	Western Europe
5	1970	1	0	NULL	0	NULL	101	Japan	4	East Asia
6	1970	1	1	NULL	0	NULL	217	United States	1	North America
7	1970	1	2	NULL	0	NULL	218	Uruguay	3	South America
8	1970	1	2	NULL	0	NULL	217	United States	1	North America
9	1970	1	10	NULL	0	NULL	499	East Germany (GDR)	9	Eastern Europe

Summary



Data Profiling



Having moved the data to MYSQL, we ran scripts that determined which columns were sparse

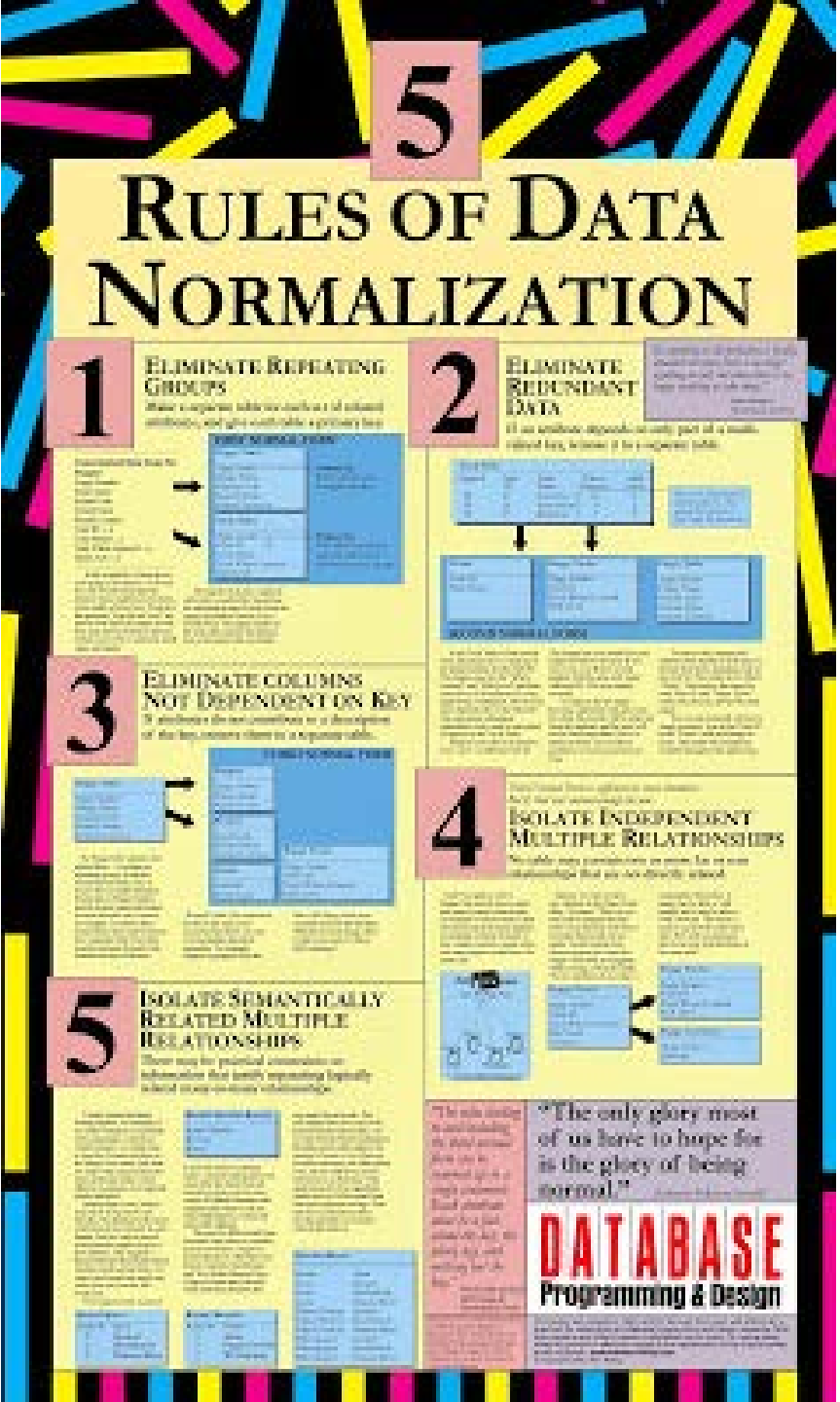
The criterion for sparseness was that 80% or more of the column was NULL or contained '.'

"When you look into an abyss, the abyss also looks into you"

-Frederick Nietzsche

We created alter tables for the sparse columns and dropped them

At this point, the table was ready for normalization

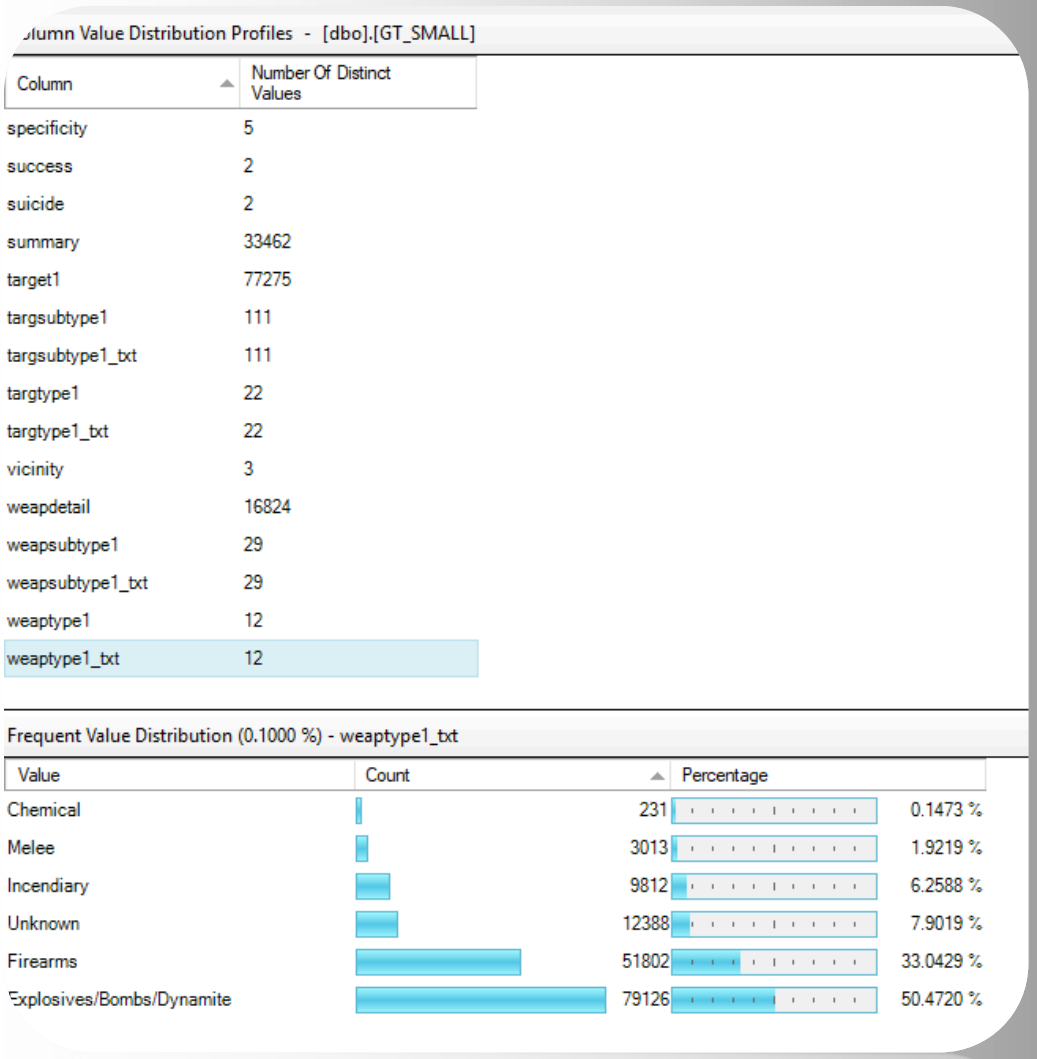


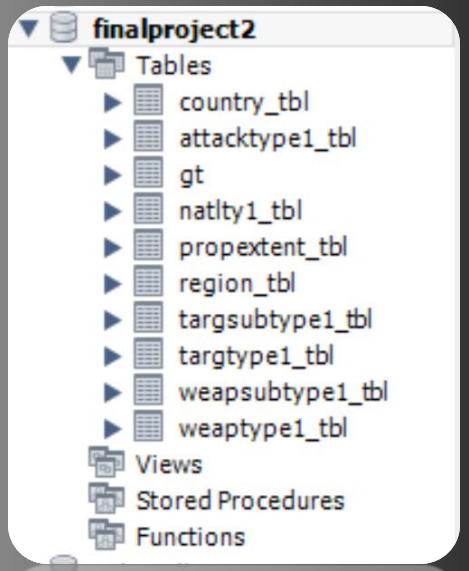
Data Normalization

- The data was profiled using SQL Server Data Profiler
- The profiler suggested that nine tables could be extracted from the large GT table.
- Individual tables were created using the columns from GT. These nine individual tables were validated by joining back to GT.
- The redundant columns were dropped form GT because they had been moved to the individual tables.
- Primary keys were created on the nine tables.
- From there, indices were created on the join columns in the GT table. Finally, a nine-way inner join was performed with the GT table to validate that everything has worked correctly.

SQL Data Profiler

Normalized Data





Data Analysis: SQL

```
SELECT country_txt, count(country_tx
FROM finalproject2.gt
group by country_txt
order by number desc;
```

country_txt	Number
Iraq	18770
Pakistan	12768
India	9940
Afghanistan	9690
Colombia	8077
Peru	6085
Philippines	5576
El Salvador	5320
United Kingdom	4992
Turkey	3557
Thailand	3338
Spain	3239
Sri Lanka	2982
Somalia	2890
Nigeria	2888
Algeria	2720
United States	2693
France	2617
Yemen	2598
Lebanon	2413
Chile	2334
Russia	2104
Israel	2085
Guatemala	2050
West Bank an...	1990

```
SELECT country_txt, count(country_txt) as
FROM finalproject2.gt
where attacktype1 = 3
group by country_txt
order by number desc;
```

country_txt	Number	attacktype1_txt
Iraq	13635	Bombing/Explosion
Pakistan	6788	Bombing/Explosion
Afghanistan	5064	Bombing/Explosion
India	4036	Bombing/Explosion
Peru	3114	Bombing/Explosion
Colombia	3036	Bombing/Explosion
France	2086	Bombing/Explosion
El Salvador	2055	Bombing/Explosion
United Kingdom	2021	Bombing/Explosion
Spain	1901	Bombing/Explosion
Philippines	1843	Bombing/Explosion
Chile	1770	Bombing/Explosion
Thailand	1538	Bombing/Explosion
Turkey	1510	Bombing/Explosion
Lebanon	1500	Bombing/Explosion
Israel	1486	Bombing/Explosion
United States	1369	Bombing/Explosion
Sri Lanka	1149	Bombing/Explosion
Russia	1124	Bombing/Explosion
Somalia	1082	Bombing/Explosion
Algeria	1069	Bombing/Explosion
Yemen	1011	Bombing/Explosion
Syria	985	Bombing/Explosion
South Africa	914	Bombing/Explosion
Ukraine	886	Bombing/Explosion
Egypt	813	Bombing/Explosion
Nigeria	786	Bombing/Explosion
Bangladesh	754	Bombing/Explosion

```
SELECT country_txt, count(country_txt) as Number
FROM finalproject2.gt
where iyear > 1970 and
iyear < 1990
group by country_txt
order by number desc;
```

country_txt	Number
El Salvador	4568
Peru	4242
Colombia	3369
United Kingdom	2940
Spain	2237
Chile	1816
Nicaragua	1770
Guatemala	1600
United States	1520
Lebanon	1347
France	1329
Sri Lanka	1323
Italy	1318
India	1252
South Africa	1041
Philippines	1032
Turkey	812
Israel	633
Argentina	605
West Germany...	512
Iran	423
Greece	423
West Bank an...	343
Pakistan	212
Bolivia	212
Honduras	202

```
SELECT attacktype1_tx
FROM finalproject2.gt
where iyear > 2007 an
iyear < 2017
group by attacktype1
order by number desc;
```

attacktype1_txt	Number
Bombing/Explosion	38424
Armed Assault	17196
Hostage Taking (Kidnapping)	4973
Assassination	4290
Facility/Infrastructure Attack	3463
Unknown	2544
Unarmed Assault	293
Hostage Taking (Barricade Incident)	180
Hijacking	167

```
SELECT attacktype1_txt, count(a
FROM finalproject2.gt
where iyear > 1970 and
iyear < 1990
group by attacktype1_txt
order by number desc;
```

attacktype1_txt	Number
Bombing/Explosion	18118
Armed Assault	9064
Assassination	7071
Facility/Infrastructure Attack	2484
Hostage Taking (Kidnapping)	1616
Unknown	1265
Hostage Taking (Barricade Incident)	499
Hijacking	162
Unarmed Assault	66

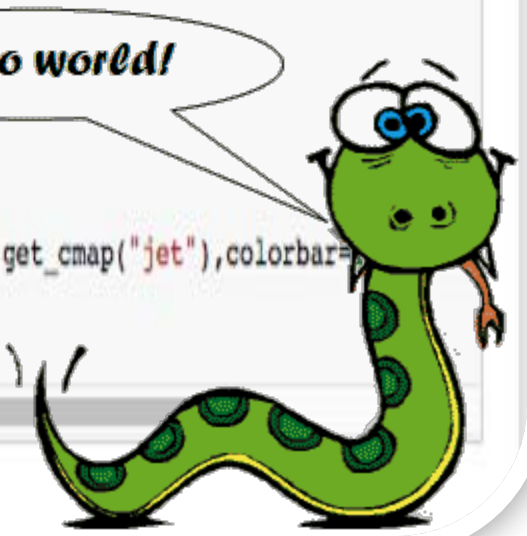
Data Analysis: Python

We used python to analyze the data, and found that weapon of choice in indigent countries is the motor vehicle.

Deduction: The favorite target of terrorists are abortion related centers and airports.

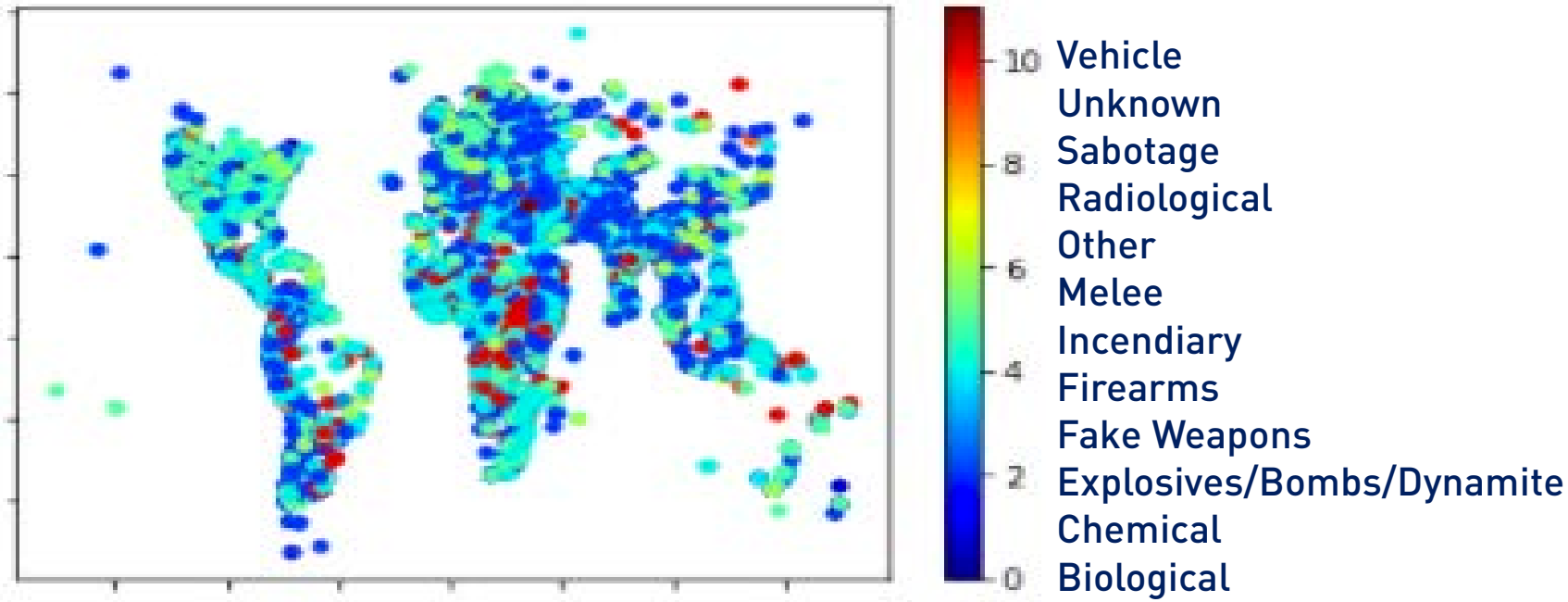
In [23]:
```
# Encode the categorical variables to see what weapons are used where
# Biological = 0, Chemical =1, etc.
# The Chosen Visualization is important because it displays what weapons are predominantly used where
# For example, in poorer countries or countries with restriction on firearms cars (denoted in red) will be used
%matplotlib inline
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
weapons = GT['weaptype1_txt']
weapons_encoded= encoder.fit_transform(weapons)
weapons_encoded
ax = GT.plot(kind='scatter', x='longitude', y='latitude', c=weapons_encoded,cmap=plt.get_cmap("jet"),colorbar=
plt.legend(bbox_to_anchor=(1,1),bbox_transform=plt.gcf().transFigure)
```

Out[23]:
```
<matplotlib.legend.Legend at 0x127bc5780>
```



Weapon Usage

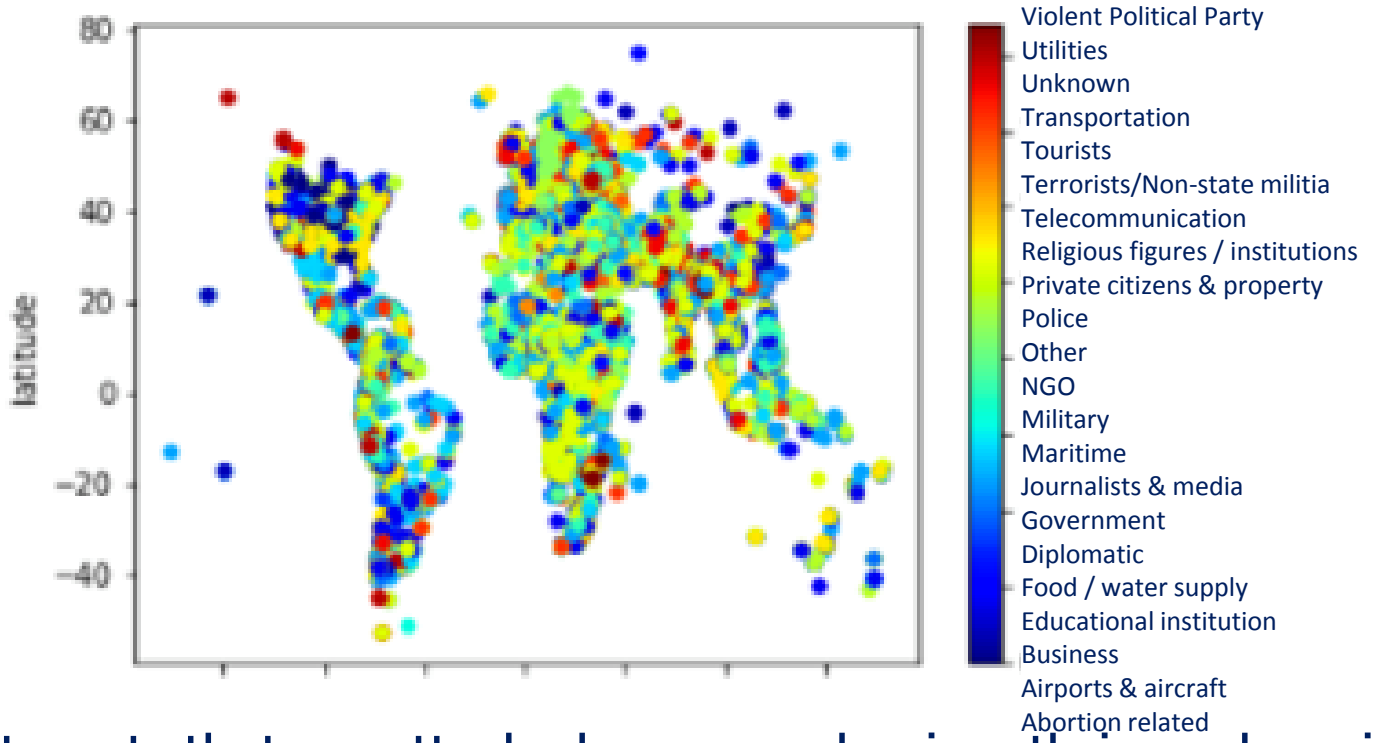
python analysis



- Encode categorical variables numerically
- Identify which weapons are used more predominantly in each region
- Developing regions : heavy use of motor vehicles as weapon of choice
- Developed nations : wide range of weapons available
  - primary methods used rely on biological, chemical, explosive devices

- **Developing regions – weapons of opportunity**
- **Developed nations – weapons of sophistication, planning, and significant impact**

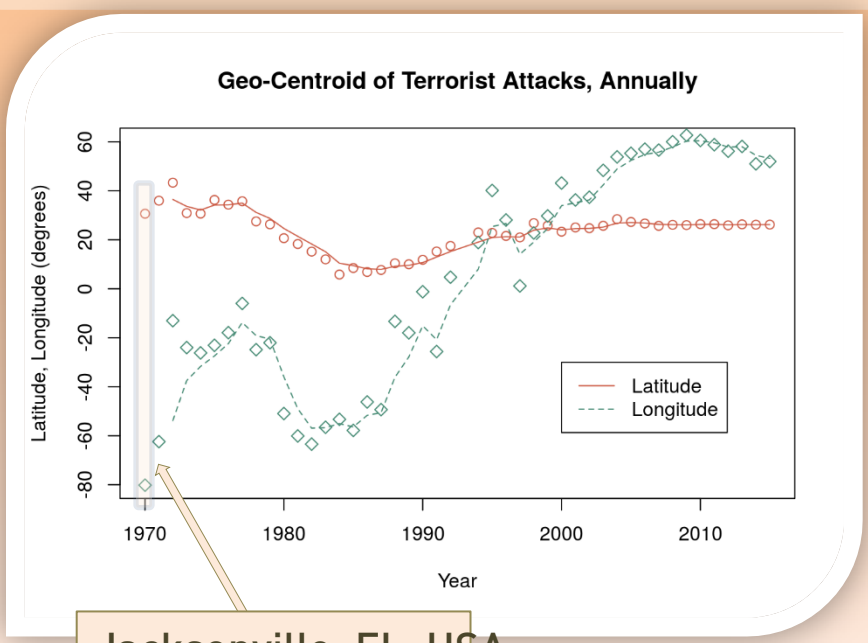
Targets

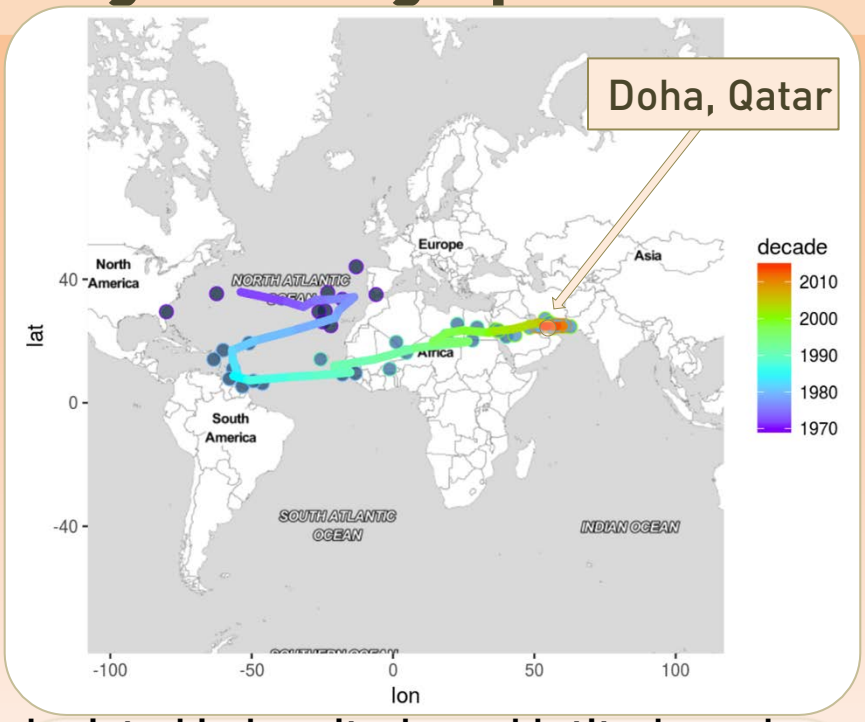


- Identify the targets that are attacked more predominantly in each region
- Violent political parties are rarely targeted
- Airports and abortion centers are often targets of terrorist activity

- **Developing regions**
  - **targets highly varied : religious, private property, tourists, transportation**
  - **targets of opportunity and localized gain**
- **Developed nations**
  - **targets of significant impact (aircraft) & specific political agendas (abortion)**
  - **targets of strategic advancement**

Data Analysis: R -Event Clustering and Geographic Trends



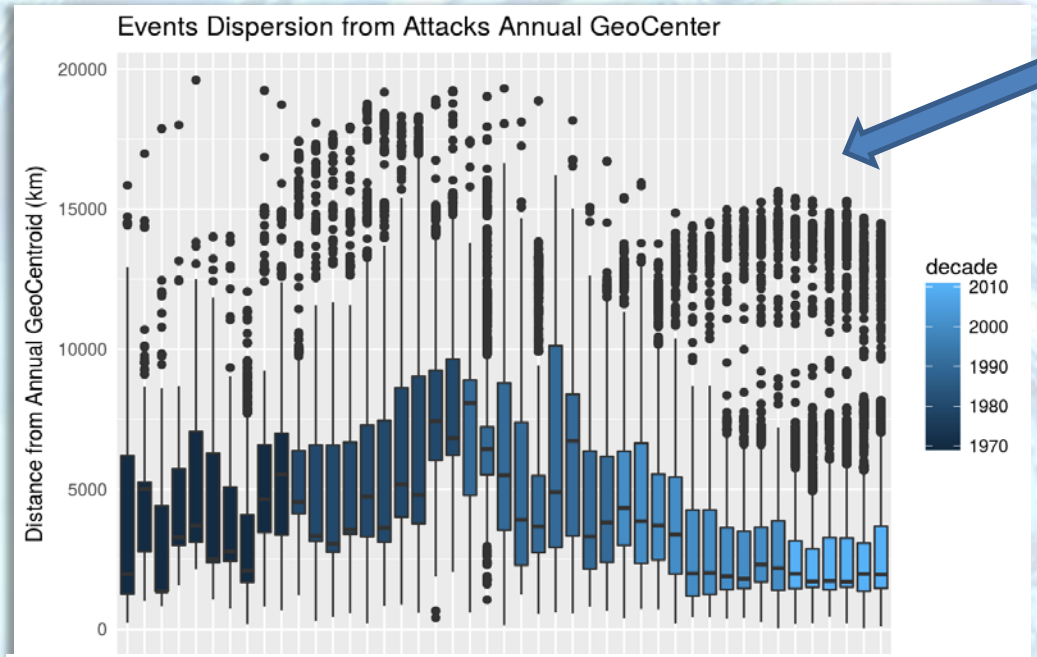


- Geographic center of attacks for each year is depicted in longitude and latitude and also represented on earth map
- *1970 - center of terrorist events was just off coast of Florida, USA*
- *each succeeding year and decade identifies migration of activity*
- *1980s - towards western Europe, then west again to northern South America*
- *1990s and after - steady migration east, to central Africa and currently in Mid-East*

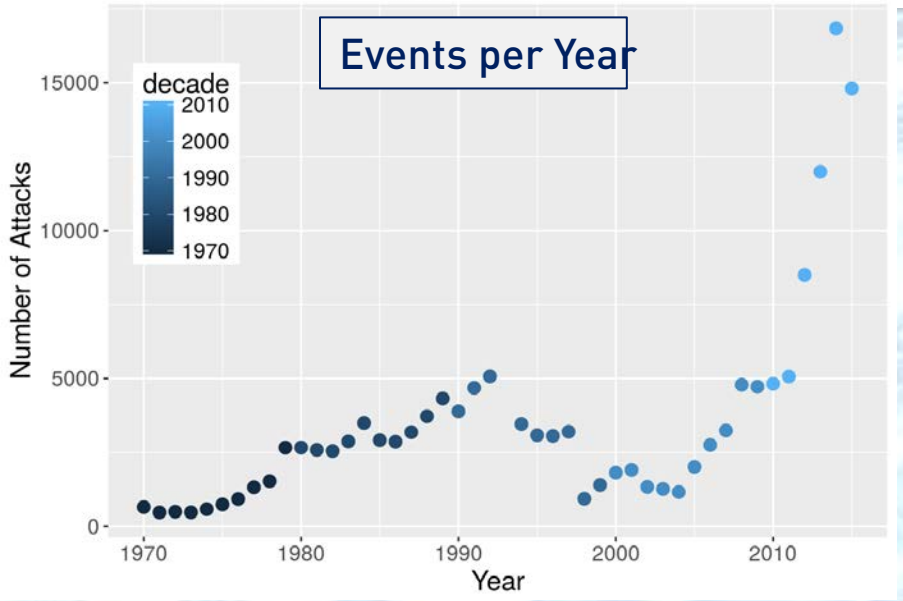
Terrorist activity – continually counter-responding to changing geo-political pressures

Current locus of activity in Mid-East, southern Asia is also likely transitory

Geographic Center and Dispersion



Distributions of distances of attacks from the annual geo-centroid



1970s - unimodal distribution, within 5000 km of geo-centroid

1980s - centroid western Europe or northern South America

distance increases slightly

bi-modal trend begins to emerge - events consistently at 5000 km from annual center

1990s - centroid begins to march eastward across Africa

contraction in population distance from the annual meanr

return to a primarily unimodal distribution, as the overall

2000s - strong concentration to the annual center - within 4000 km of the annual center

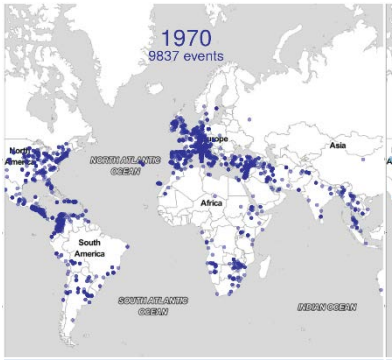
reappearance of a bi-modal distribution

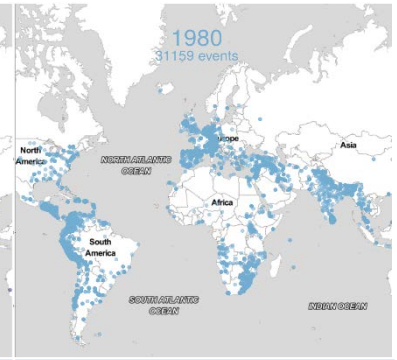
2010s - least dispersion in the overall population difference from the annual centers.

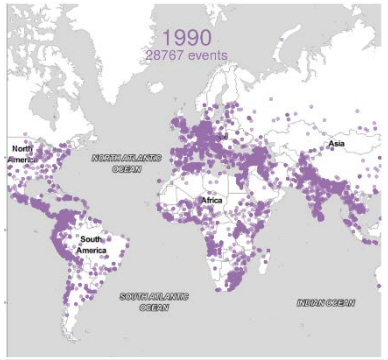
Thus, in the current era, there is a very strong concentration of terrorist events in the Mid-East, and also two other locations affected by ~25% of the events, as shown by tri-modal distribution.

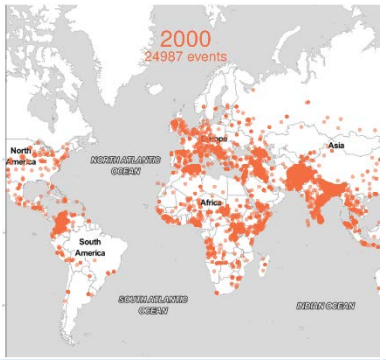
Lexicon of terrorism

Text mine : frequent and novel words, by decade









1970s	1980s	1990s	2000s	2010s
vietnam	ashley	gia (Algeria)	Kandahar, khost	asimah, huthis
selective	beseda	angola	basra, diyala, mosul	donetsk, luhansk
r.o.t.c	brockhoeft	india	ninawa	homs
draft	toledo	ireland	kirkuk	shebelle
colorado	contras	kosovo	khan	amanat, pakhtunkhwa
iowa	nicaragua	pakistan	aqlim	borno
manhattan	spinks	pastrana (Colombia)	pattani, yala	derna, sirte
ohio	sinhalese	samper (Colombia)	al qaida	hartal, ips, sabha
mexican	ltte		gaza	zuweid
puerto rico	csa	extremists	shi'a	haram, maiduguri
firebombed			somali	aqap
				• saladin

Viet Nam War
USA sites
Puerto Rican independence

Anti-abortion
Anti-communism
Munich Olympics
Sri Lanka
White supremacy

5 continents
terrorism is global

1st use of extremist as frequent description

Afghanistan
Iraq
Iran
Kurdistan
London
Maghreb
Thailand

Yemen, Ukraine
Syria, Somalia
Pakistan, Nigeria
Libya, India
Egypt, Boko Harum
• 12th century

Analysis Insights

- Developing vs. Developed Regions
  - Weapons comparison
    - opportune vs. sophisticated and larger impact
  - Targets comparison
    - opportune & localized gain vs. strategic
- Geographic Trends
  - N. America → Europe → S. America – Africa – Mid-East / S. Asia
  - Dispersion and distribution – non-constant, currently heavily focused in S. Asia, with significant percentage of events in Africa and Mid-East
  - Exponential growth in last 5 years (3x)
- Lexicon – Text Mining
  - Simple question : what words are new in event summary reporting ?
  - Provides opportunity for historical context
  - Ever evolving nature of global terrorism – causes, effects, locations

Summary

- GTD – open source data
- Raw data required significant manipulation
  - wrangled, organized, sql-ized, cleaned, normalized
- Produced MySQL version of dataset for team exploitation
- Analysis insights produced
  - Developing vs. Developed regions
  - Geographic trends
  - Lexical trends

Recent history of terrorism : non-constant
Exploring evolution of modern terrorism provides context of current events and prepares us for future developments