

Introduction	Data sources 000	Data exploration	Results 000 000	Conclusion	Files

Table of contents I



- 2 Data sources
- 3 Data exploration









Introduction	Data sources 000	Data exploration	Results 000 000	Conclusion	Files

A quick overview

Questions that were/are interesting:

- What is the GDELT project
- What data is available
- What does the data have big data potential



◆□> <畳> <目> <目> <目> <<=>

Introduction	Data sources ●○○	Data exploration	Results 000 000	Conclusion	Files
Data from multiple	e sources				
	T claims th	o following:			

GDELT claims the following:

- Real-time Translation of 65 Languages.
- **2** Real-time Measurement of 2,300 Emotions and Themes.

<□ > < @ > < ≧ > < ≧ > ≧ > < ○ 4/15

- **③** High Resolution View of the Non-Western World.
- Relevant Imagery, Videos, and Social Embeds.
- **9** Quotes, Names, and Amounts.

Introduction	Data sources ○●○	Data exploration	Results 000 000	Conclusion	Files
Data from multiple	sources				

GDELT data is available on Google BigQuery

Basically:

- You must have a Google account.
- The first 1TB of data processed per month is free.
- After the first 1TB, you are charged for the number of bytes processed, or stored.
 - Processed: \$5 per TB
 - Storage: \$0.02 per GB per month
- Costs above the free tier are charged to your credit card on file with Google.

• • • REALIZED BEEN	a contraining contrainty	ever(6)		* C Stach 4 6 0 4 1
Google BigQuery				o 😏
controls quality	Nex-Query (1)			Quey Linter 100 Linter 1
Query History 200 History				8
Relytorial (8)				
Ay Propert R				
No-below function for project				
Place-orde a didance tabula ten procedurative nere above.	NAMES OF	ter (ten	loss the	Permit Overy ReserVapleer
passag .				
a sea	Table Details: e	reals		Grow Table Cross Table Expect Table Device Table
* 14	Table Details: e	Protor		Soverheim Das Teller Band Teller
v M II. result, prostprotes II. nees		Rotor	NULLARS	Dear Service Dear
V M II recent, provedpresses II recent II recent pattornel	Scheres Dotato	Protor		
v M III menak_possetynako III nem	bhana brana BLOBHLEMBATD	Poster N°EXIP	HULLMER	Unget & treatment
v tel i creenak_posserpusket i nem justonel > polst polst_puen. > telminations	bown bown BOBALT	Poter N'ESO N'ESO	NULLMER	VNgA D fratheet De Breathe plat / VVVMDO beat
v ká i rezená, poszemprokok, i neve i nevez preson k pánki k pánki pyen.	Science Docks Robert Docks Robert Re- Robert Re- Tear	Poster PCEDD PCEDD PCEDD PCEDD	NULLARE NULLARE NULLARE	Annual (Marchan)
V VAL III reeses possesponents III reeses III reese	bhans brads B.OBM.BVBWPD BQCMPE Manishing	Poter N'ESO N'ESO	NULLMER	Annual (Marchan)
Vial Proving percentageological Percentageo	Science Docks Robert Docks Robert Re- Robert Re- Tear	Poster PCEDD PCEDD PCEDD PCEDD	NULLARE NULLARE NULLARE	And Sha

Introduction	Data sources ○○●	Data exploration	Results 000 000	Conclusion	Files
Data from multiple	e sources				

It is easy to blow through 1TB

```
SELECT DATE, coord, cnt from (
SELECT DATE, coord, COUNT(*) as cnt
FROM (
select date, REGEXP_REPLACE(REGEXP_EXTRACT(
SPLIT(V2Locations,';')
,r'^[2-5]#.*?#.*?#.*?#(.*?#.*?)#')
, '^{(.*?)}#(.*?)', '^{1}' as coord
from [gdelt-bq:gdeltv2.gkg]
where DATE >= 20150322000000 and DATE <= 20150328999999 )
where coord is not null group by date, coord ORDER BY 3 DE
where cnt >= 3;
```

The script takes about 15 seconds to run, and processes 159GB of data.

Introduction	Data sources 000	Data exploration	Results 000 000	Conclusion	Files

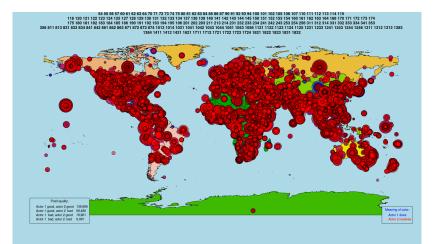
About the GDELT data

- Actors 1 and 2. Actor 1 does something to Actor 2. Each actor is identified by lots of fields.
- Every event has action attributes. Every event has lots of fields.
- Geographic information about actor 1 and actor 2. Data beyond position.

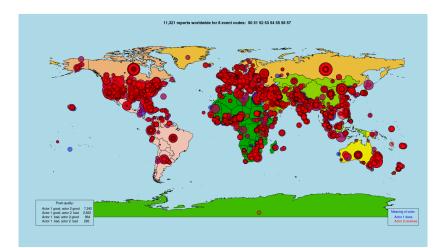
Our explorations will focus on a single day 17 October 2017.

Introduction	Data sources 000	Data exploration	Results •00 000	Conclusion	Files
Graphical data					

All the events in one day



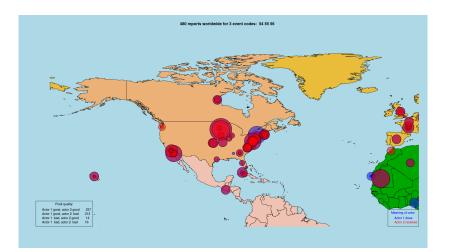
Introduction	Data sources 000	Data exploration	Results ○●○ ○○○	Conclusion	Files
Graphical data					
Event	codes 50 -	57			



Introduction	Data sources	Data exploration	Results	Conclusion	Files
	000		000		

Graphical data

Event codes 54 - 56 for the North West quadrasphere



Introduction	Data sources	Data exploration	Results	Conclusion	Files
	000		000 000		

Tabular data

The six most frequently reported event codes.

Table: The six most frequently reported event codes.

Code	Explanation	Reports
10	Make public statement	17,551
42	Consult, make a visit	16,373
43	Consult, host a visit	15,039
40	Consult	13,606
30	Express intent to cooperate	12,691
51	Praise or endorse	12,500

Introduction	Data sources 000	Data exploration	Results ○○○ ○●○	Conclusion	Files
Tabular data					

Completeness of actor codes.

Table: Completeness of actor codes.

		Actor 1	
		Good	Bad
Actor2	Good	143,294	20,272
ACLOIZ	Bad	59,874	0

Introduction	Data sources 000	Data exploration	Results ○○○ ○○●	Conclusion	Files
Tabular data					

Most common source domains.

Table: Most common source domains.

Source domain	Count
www.yahoo.com	1,502
allafrica.com	1,328
www.newsnow.in	923
www.dailymail.co.uk	900
www.4-traders.com	891
www.thehindu.com	547

<ロ > < 部 > < 言 > < 言 > こ う < つ 3/15

Introduction	Data sources 000	Data exploration	Results 000 000	Conclusion	Files

What have we covered?

- GDELT data is available on Google BigQuery site
- GDELT data is noisy
- GDELT data useful for long term trend analysis
- GDELT data can grow to Big Data size

IS THIS THE ENDP

< □ ▶ < □ ▶ < 三 ▶ < 三 ▶ 三 りへで 14/15

Next time: see where the wind blows

Introduction	Data sources 000	Data exploration	Results 000 000	Conclusion	Files
Files	of interest				
go	deltReport.R – an R s	script to	Conflict and M		
	plore the GDELT da		Observations E	vent and Actor	
			Codebook		
	AMEO.Manual.1.1b3	.par – tne	Constant		