Political Event Data

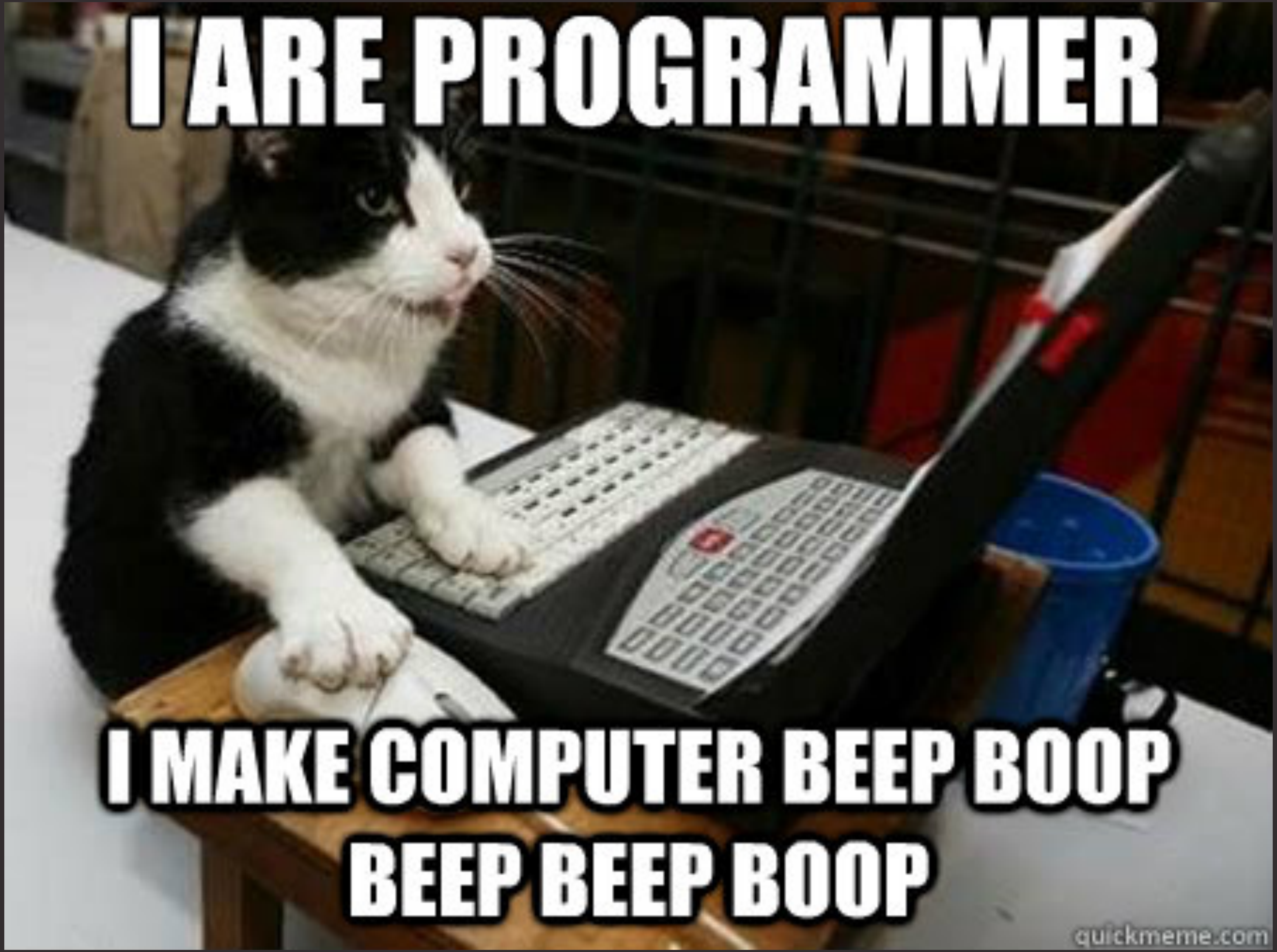
John Beieler
johnb30@gmail.com

About Me

- Political scientist (sometimes…)
- PhD candidate at Penn State
  - ABD 4 lyfe
- Work at Caerus Associates doing data things

tl;dr

A bunch of political scientists (and an undergraduate computer scientist) try to make structured data from unstructured news reporting. It works most of the time.


I ARE PROGRAMMER
I MAKE COMPUTER BEEP BOOP
BEEP BEEP BOOP
quickmeme.com

What?

Who-did-what-to-whom

Who-did-what-to-whom

Syrian rebels attacked the town of Aleppo.

- Syrian rebels
- Attacked
- Aleppo

Event Codings

- Source-Action-Target
- `SYRREB 19 SYR`

Event Codings

- Dictionary-based lookups
- CAMEO coding scheme for actions
    - 20 top-level categories
    - ~240 total classifications
    - 4 (or 5…) level delineation
- Noun phrase -> code mapping

How?

The Old School

- Download a ton of text from Lexis-Nexis
    - Usually done by unhappy undergrads or graduate students
- Run it through TABARI

The Old School

TABARI

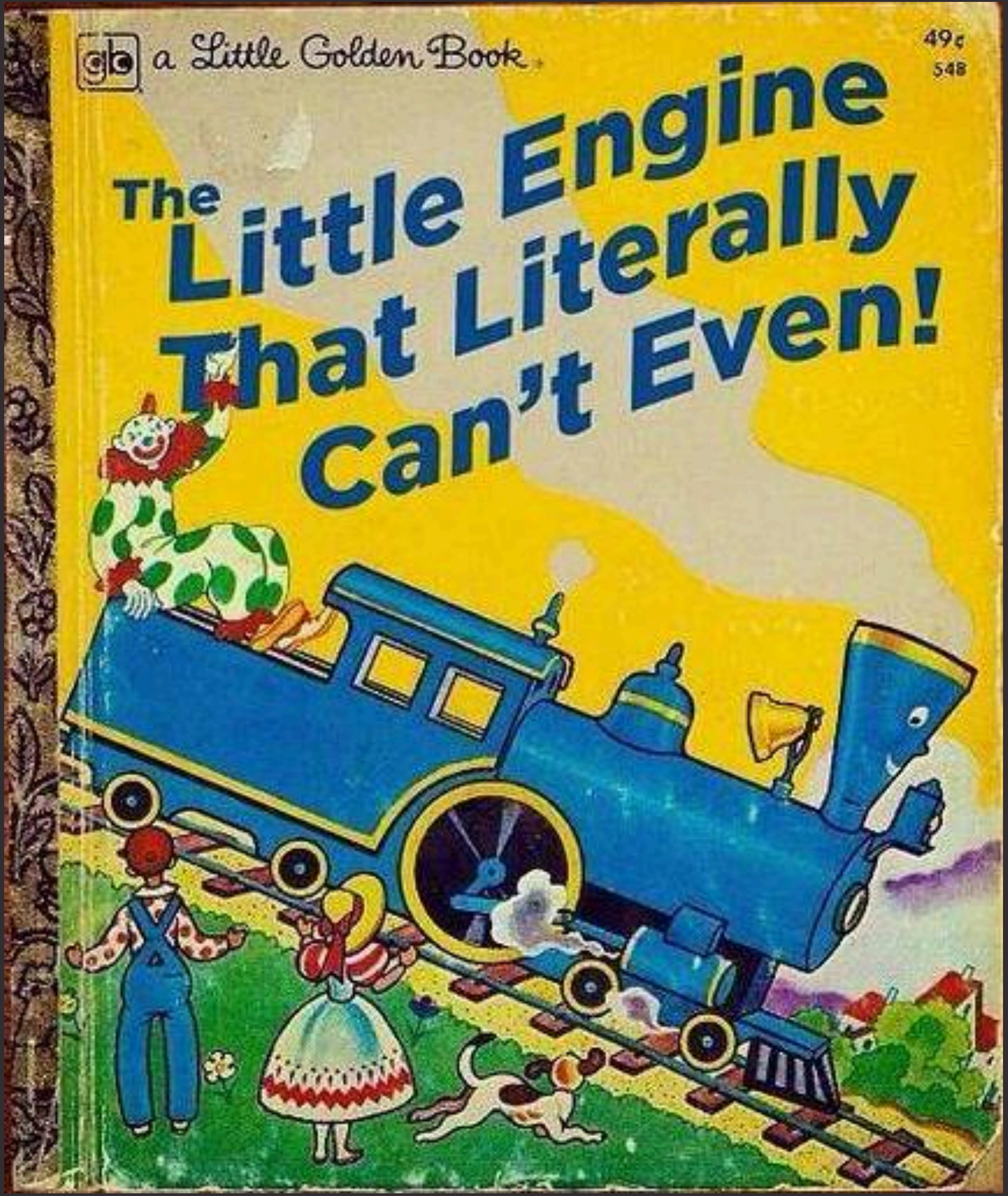
- Shallow parse
- Part-of-speech tagging
- A lookup of noun and verb phrases

Welcome to 2014

Generation 2 (3?)

- Web scraper
  - ~500 news sites
- Run it through PETRARCH

Web Scraping


gb a Little Golden Book
49¢
548
The Little Engine That Literally Can't Even!

Web scraping

- RSS feeds
- Distributed scraper
- Goose (Python)

PETRARCH

- Deep parse
- Stanford's CoreNLP
- Full parse tree
- A (more accurate) lookup of noun and verb phrases

Welcome to 2015

PETRARCH 2

The PETRARCH-aning

- Actually, really use the tree information
- Let meaning of branch phrases rise through the parse tree
  - Agent codes make more sense now
  - All kinds of cool interactions
- Reformat the underlying dictionaries

PETRARCH 2

- Read Stanford CoreNLP parse into memory using Phrase classes.
- Identify coded actors in noun phrases.
- Identify the usage of the verbs in the verb phrases based on the dictionary entries.
- Identify how verbs interact with their constituent verb, prepositional, and noun phrases.

PETRARCH 2

- Identify how verbs interact with the noun phrases in their subject position.
- Resolve verb+verb interactions.
    - "A says A attacked B" vs "A says B attacked C."
- Return the coding of the uppermost VerbPhrase, if it satisfies the conditions specified by the user

Why PETRARCH 2?

Would you like to know more?

github.com/openeventdata

github.com/caerusassociates

Why?

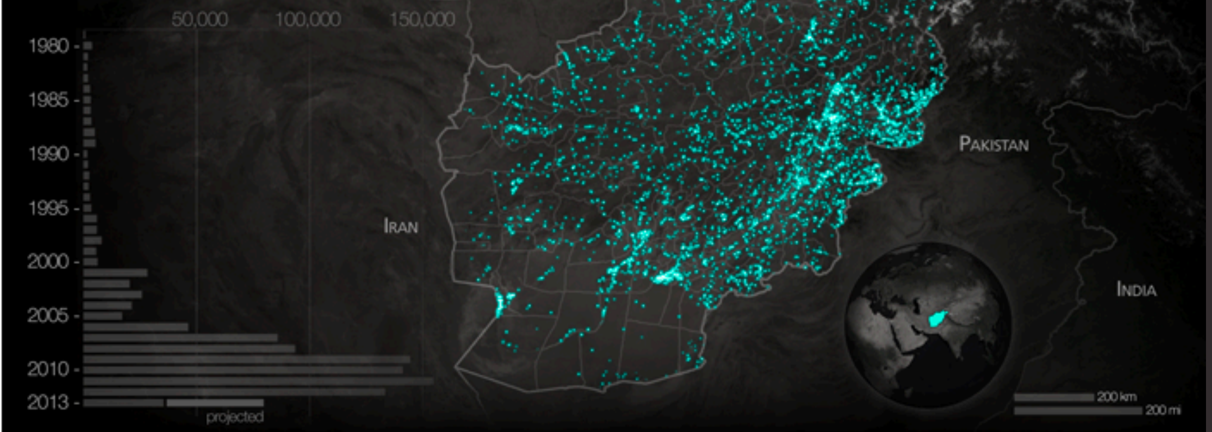
I DUNNO LOL

Many applications

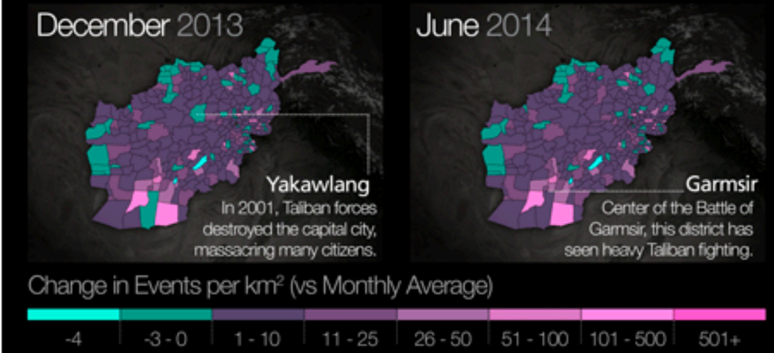
- Monitoring
- Forecasting
- Statistical inference

1,029,479 material conflicts in Afghanistan

January, 1979 - June, 2013



What the future holds: Forecasted Events



Yakawlang
In 2001, Taliban forces destroyed the capital city, massacring many citizens.

Garmsir
Center of the Battle of Garmsir, this district has seen heavy Taliban fighting.

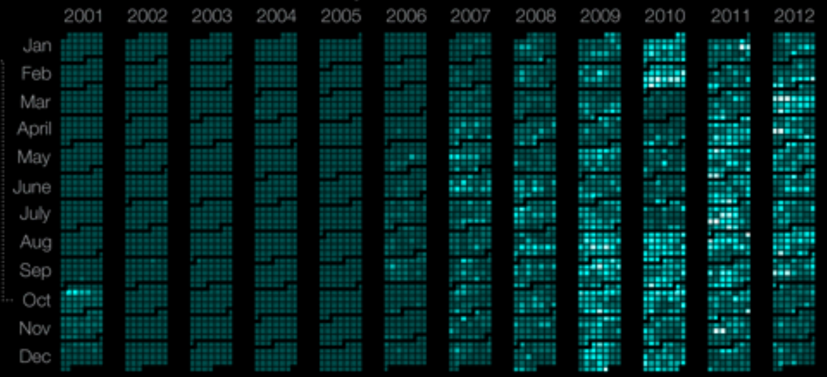
Change in Events per km² (vs Monthly Average)

Notable Regions of Conflict

Khanashin: Although the region is among the poorest in the country, its soil contains more than an estimated $85 million USD in rare earth materials and niobium. **Kandahar**: Home to the country's second largest city, the area has been the seat of many terrorist groups and targeted killings. Ghulam Haider Hamidi, former mayor of Kandahar City, was assassinated in 2011. **Kunduz:** After US forces caused the Taliban to flee Kabul, many sought refuge in Kanduz City.

Operation Enduring Freedom: Material Conflicts Per Day

The war in Afghanistan, which carries the official title *Operation Enduring Freedom*, began on **October 7, 2001**. Since that time, 948,311 material conflicts (92% of the 1979-2012 total) have occurred. More than 60 countries supported the US in the war, either directly (e.g., troops from coalition forces) or indirectly (e.g., use of airspace, facilities, or through financial backing).



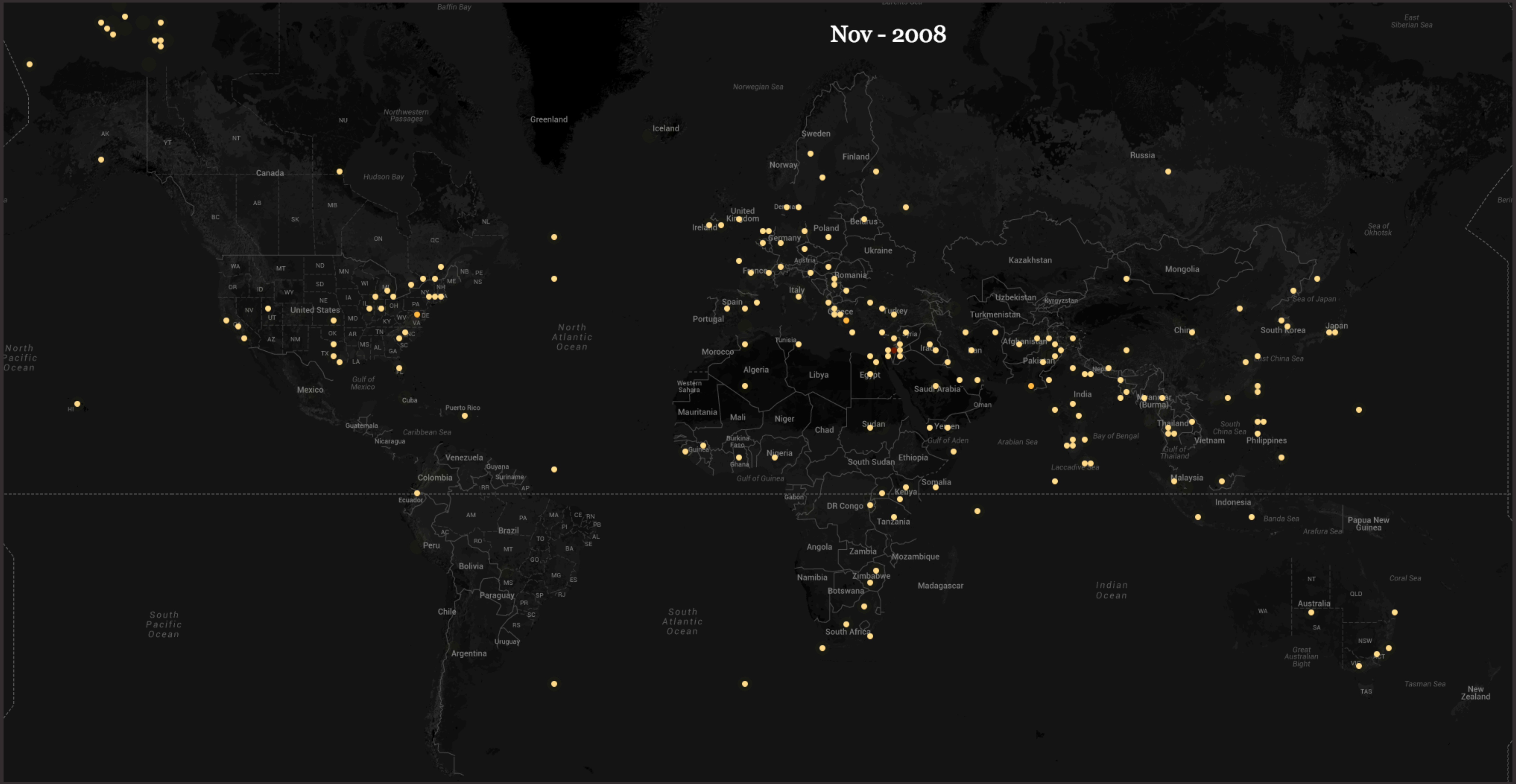
Few events → Many events

Learn more about #GDELT at: http://eventdata.psu.edu/data.dir/GDELT.html

Source: Leetaru, Kalev and Schrodt, Philip. (2013). *GDELT: Global Data on Events, Language, and Tone, 1979-2012.* International Studies Association Annual Conference, April 2013. San Diego, CA.

Maps and Visualization
Joshua Stevens @jscarto
JoshuaStevens.net

Forecasting Models and Predictions
John Beieler @johnb30
JohnBeieler.org
Jay Yonamine @jay_yonamine
JayYonamine.com



Nov - 2008

Baffin Bay
Greenland
Iceland
Norwegian Sea
Northwestern Passages
Hudson Bay
Canada
United States
Mexico
Cuba
Puerto Rico
Caribbean Sea
Gulf of Mexico
Guatemala
Nicaragua
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Argentina
Uruguay
North Pacific Ocean
South Pacific Ocean
North Atlantic Ocean
South Atlantic Ocean
Sweden
Finland
Norway
United Kingdom
Ireland
Denmark
Poland
Belarus
Germany
Ukraine
Austria
France
Romania
Italy
Spain
Portugal
Greece
Turkey
Tunisia
Morocco
Algeria
Libya
Egypt
Western Sahara
Mauritania
Mali
Niger
Chad
Sudan
Burkina Faso
Guinea
Ghana
Nigeria
Gulf of Guinea
Gabon
South Sudan
Ethiopia
Somalia
Kenya
DR Congo
Tanzania
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
South Africa
Madagascar
Saudi Arabia
Oman
Yemen
Gulf of Aden
Iraq
Iran
Russia
Kazakhstan
Uzbekistan
Kyrgyzstan
Turkmenistan
Afghanistan
Pakistan
Nepal
India
Arabian Sea
Laccadive Sea
Bay of Bengal
Indian Ocean
Mongolia
China
South Korea
Japan
Sea of Japan
Sea of Okhotsk
East Siberian Sea
East China Sea
Myanmar (Burma)
Thailand
Gulf of Thailand
Vietnam
South China Sea
Philippines
Malaysia
Indonesia
Banda Sea
Arafura Sea
Papua New Guinea
Coral Sea
Australia
Great Australian Bight
Tasman Sea
New Zealand

Context

Problems

Oh

The

Problems

Problems

- Clean text?
- Relevant stories?
- New actors?
- New action categories?
- Error propagation.
- How good is any of this?

BRACE YOURSELVES
THE AREA SPECIALISTS ARE COMING
imgflip.com

Geolocation

Geolocation

- Could be a whole talk
- Many solutions
  - CLAVIN
  - CLIFF
  - Mordecai
- None perfectly suited to this application

Future

Future

- Something other than CAMEO
- Better document ingest
- Geolocation

GDELT

Global Database of Events, Language, and Tone

ICEWS

Integrated Crisis Early Warning System

Phoenix

No longer "one dataset to rule them all"

Questions?

We're hiring.

caerusassociates.com/careers