# GDELT: Global Data on Events, Location and Tone

Philip A. Schrodt

Parus Analytical Systems
schrodt735@gmail.com

Workshop at the Conflict Research Society
Essex University
17 September 2013

# Overview

- What is GDELT
  - 250M geolocated events, 1979-present
  - Updated daily: `gdelt.utdallas.edu`
  - Extended user community: `gdeltbog.wordpress.com`, `#gdelt`
- GDELT 1.0 Framework
  - Variety of news sources which change significantly over time
  - TABARI coding engine with customized geolocation software
  - CountryInfo.txt dictionaries
- Difficulties and limitations
  - Sheer size of the dataset
  - Base level of events increases exponentially after 2002
  - Very high number of false positives
- GDELT 2.0 Enhancements
  - PETRARCH and Stanford CoreNLP coding
  - Google Translate input
  - WordNet and NER-enhanced dictionaries
  - Additional community enhancements?

# Topics:

Overview

The Big Picture

Visualizations

Why GDELT is difficult to use

GDELT 2.0

Named Entity Recognition/Resolution

Additional work to be done

# Basics

- Coverage: 1979-present with daily updates
- Size: 250-million events
- Coding System: CAMEO
- Coding Engine: TABARI 0.8 for actors and events; custom software for geolocation
- Geolocation: separate fields for source, actor and event; resolved to city level
- License: open source for both data and software

# News Story Example: Example: 18 December 2007

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

The Turkish attacks in Dohuk Province on Sunday—involving dozens of warplanes and artillery—were the largest known cross-border attack since 2003. They occurred with at least tacit approval from American officials. The Iraqi government, however, said it had not been consulted or informed about the attacks.

Massoud Barzani, leader of the autonomous Kurdish region in the north, condemned the assaults as a violation of Iraqi sovereignty that had undermined months of diplomacy. “These attacks hinder the political efforts exerted to find a peaceful solution based on mutual respect.”

New York Times, 18 December 2007
http://www.nytimes.com/2007/12/18/world/middleeast/18iraq.html?_r=1&ref=world&oref=slogin (Accessed 18 December 2007)

# TABARI Coding: Lead sentence

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: First event

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Actors

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Agent

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Second event

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Second event target

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Agent

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# Mommy, where does GDELT come from?

# Mommy, where does GDELT come from?

# GDELT Timeline

| | |
|---|---|
| Fall 2011 | Kalev downloads stories, begins web scraping and negotiating access to bulk feeds |
| 2012 | |
| Spring | TABARI used to code 1979-2012 |
| August | Prototype provided to PSU |
| September | First GDELT Hackathon |
| 2013 | |
| March | Static GDELT released on PSU site in conduction with ISA |
| April | Syria, Afghanistan graphics in *Guardian, Foreign Policy* |
| June | UT/Dallas server operational with daily updates |
| July | gdeltblog.wordpress.com |
| August | Beiler/Stevens protest graphic receives 150,000+ views, Chelsea Clinton tweet |

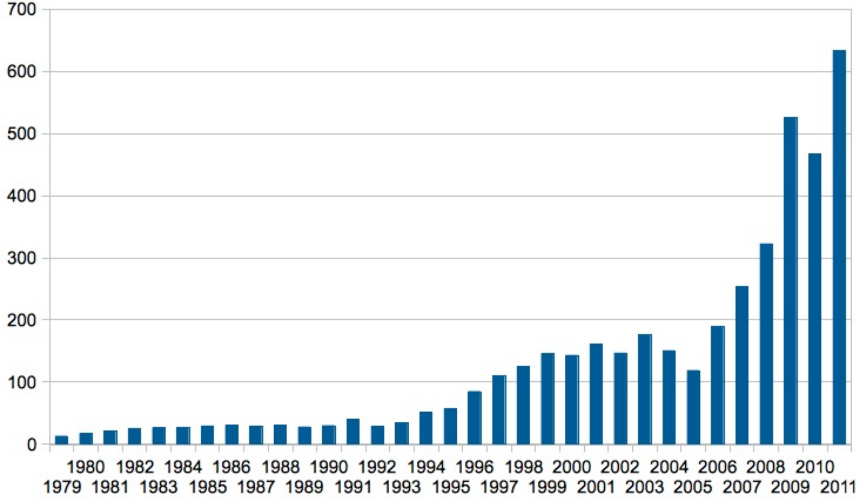
# GDELT Community

- Blog: http://gdeltblog.wordpress.com
- Twitter: #gdelt
- Github: eventdata
- Collective development of tools, mostly in R and in geospatial data analytics

# Sources

- 1979-present: Agence France Press, Associated Press, Xinhua
- 2000-present: BBC Monitor
- 2002 (?) - present: Google News
- 2013+(?): Google Translate

# Density of Data across Time (Gb per year)



# Processing Pipeline

- Downloading and formatting stories
- Pre-processing entity names to make these “TABARI-friendly”:

  essentially named-entity-resolution (NER)
- TABARI coding
- Separate coding for geolocation using customized software: this should be available on the web in the very near future. Note that geocoding is generally only to the city level, and in many cases resolves to the country level, where it is assigned the country centroid.

# Documentation

- Event data generally: Schrodt and Gerner 2000/2012 *Analyzing International Event Data*, chapts 1-3, `http://eventdata.parusanalytics.com/papers.dir/automated.html`.
- Current formats: `http://gdelt.utdallas.edu`
- Current tools: `http://gdeltblog.wordpress.com`

We are also hoping to get an general textbook going using an open-collaboration environment: this will cover both event data analysis and the toolset.

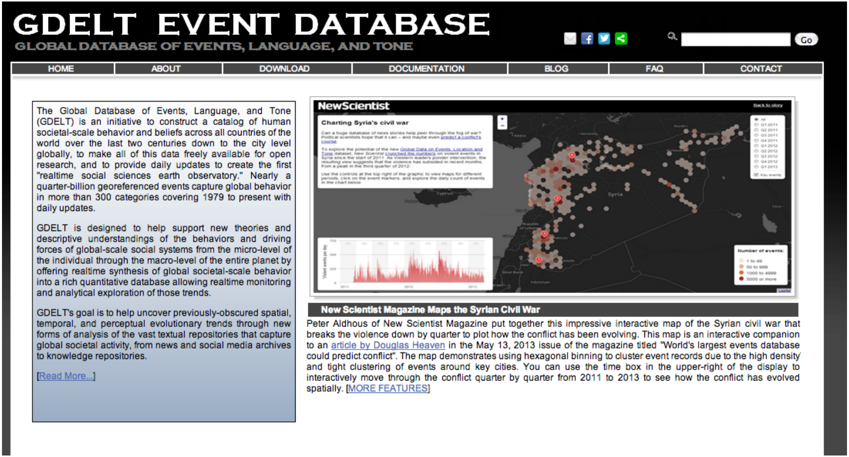
# Textual Analysis By Augmented Replacement Instructions (TABARI)

- ANSI C++, approximately 14,000 lines of code
- Open-source (GPL)
- Unix, Linux and OS-X operating systems (gcc compiler)
- “Teletype” interface: text and keyboard
  - Easily deployed on a server
- Codes around 5,000 events per second on contemporary hardware
  - Speed is achieved through use of shallow parsing algorithms
  - Speed can be scaled indefinitely using parallel processing
- Standard dictionaries are open source, with around 15,000 verb phrases for events and 30,000+ noun phrases for actors
- Coded the entire GDELT dataset without crashing

# CAMEO

- 20 primary event categories; around 200 subcategories
- Based on the WEIS typology but with greater detail on violence and mediation
- Combines ambiguous WEIS categories such as [WARN/THREATEN] and [GRANT/PROMISE]
- National actor codes based on ISO-3166 and `CountryInfo.txt`
- Substate "agents" such as GOV, MIL, REB, BUS
- Extensive IGO/NGO list

# http://gdelt.utdallas.edu



The Global Database of Events, Language, and Tone (GDELT) is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world over the last two centuries down to the city level globally, to make all of this data freely available for open research, and to provide daily updates to create the first "realtime social sciences earth observatory." Nearly a quarter-billion georeferenced events capture global behavior in more than 300 categories covering 1979 to present with daily updates.

GDELT is designed to help support new theories and descriptive understandings of the behaviors and driving forces of global-scale social systems from the micro-level of the individual through the macro-level of the entire planet by offering realtime synthesis of global societal-scale behavior into a rich quantitative database allowing realtime monitoring and analytical exploration of those trends.

GDELT's goal is to help uncover previously-obscured spatial, temporal, and perceptual evolutionary trends through new forms of analysis of the vast textual repositories that capture global societal activity, from news and social media archives to knowledge repositories.

[Read More...]

**Charting Syria's civil war**

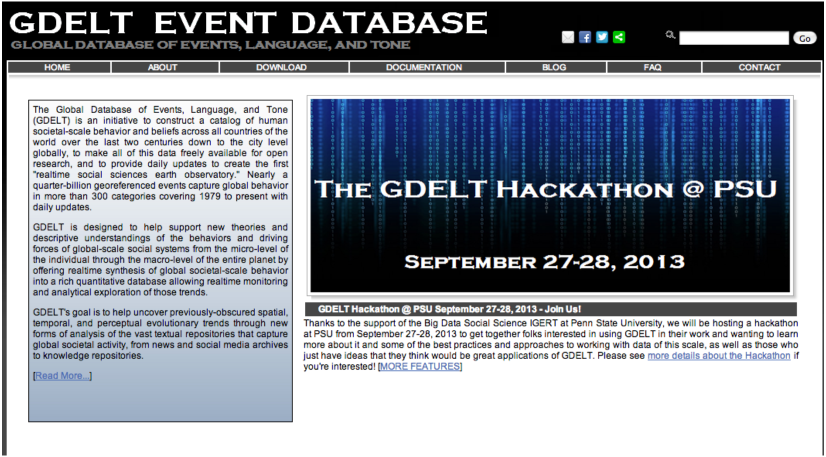
Can a huge database of news stories help peer through the fog of war? Political scientists hope that it can – and maybe even predict a conflict's course.

To explore the potential of the new Global Data on Events, Location and Tone dataset, *New Scientist* crunched the numbers on violent events in Syria since the start of 2011. As Western leaders ponder intervention, the resulting view suggests that the violence has subsided in recent months, from a peak in the third quarter of 2012.

Use the controls at the top right of the graphic to view maps for different periods, click on the event markers, and explore the daily count of events in the chart below.

**New Scientist Magazine Maps the Syrian Civil War**

Peter Aldhous of New Scientist Magazine put together this impressive interactive map of the Syrian civil war that breaks the violence down by quarter to plot how the conflict has been evolving. This map is an interactive companion to an article by Douglas Heaven in the May 13, 2013 issue of the magazine titled "World's largest events database could predict conflict". The map demonstrates using hexagonal binning to cluster event records due to the high density and tight clustering of events around key cities. You can use the time box in the upper-right of the display to interactively move through the conflict quarter by quarter from 2011 to 2013 to see how the conflict has evolved spatially. [MORE FEATURES] http://gdelt.utdallas.edu



The Global Database of Events, Language, and Tone (GDELT) is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world over the last two centuries down to the city level globally, to make all of this data freely available for open research, and to provide daily updates to create the first "realtime social sciences earth observatory." Nearly a quarter-billion georeferenced events capture global behavior in more than 300 categories covering 1979 to present with daily updates.

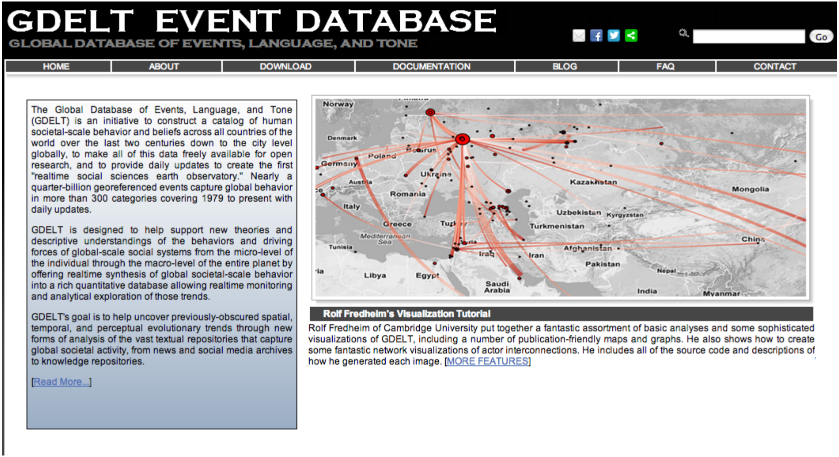
GDELT is designed to help support new theories and descriptive understandings of the behaviors and driving forces of global-scale social systems from the micro-level of the individual through the macro-level of the entire planet by offering realtime synthesis of global societal-scale behavior into a rich quantitative database allowing realtime monitoring and analytical exploration of those trends.

GDELT's goal is to help uncover previously-obscured spatial, temporal, and perceptual evolutionary trends through new forms of analysis of the vast textual repositories that capture global societal activity, from news and social media archives to knowledge repositories.

[Read More...]

**GDELT Hackathon @ PSU September 27-28, 2013 - Join Us!**

Thanks to the support of the Big Data Social Science IGERT at Penn State University, we will be hosting a hackathon at PSU from September 27-28, 2013 to get together folks interested in using GDELT in their work and wanting to learn more about it and some of the best practices and approaches to working with data of this scale, as well as those who just have ideas that they think would be great applications of GDELT. Please see more details about the Hackathon if you're interested! [MORE FEATURES] http://gdelt.utdallas.edu



The Global Database of Events, Language, and Tone (GDELT) is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world over the last two centuries down to the city level globally, to make all of this data freely available for open research, and to provide daily updates to create the first "realtime social sciences earth observatory." Nearly a quarter-billion georeferenced events capture global behavior in more than 300 categories covering 1979 to present with daily updates.

GDELT is designed to help support new theories and descriptive understandings of the behaviors and driving forces of global-scale social systems from the micro-level of the individual through the macro-level of the entire planet by offering realtime synthesis of global societal-scale behavior into a rich quantitative database allowing realtime monitoring and analytical exploration of those trends.

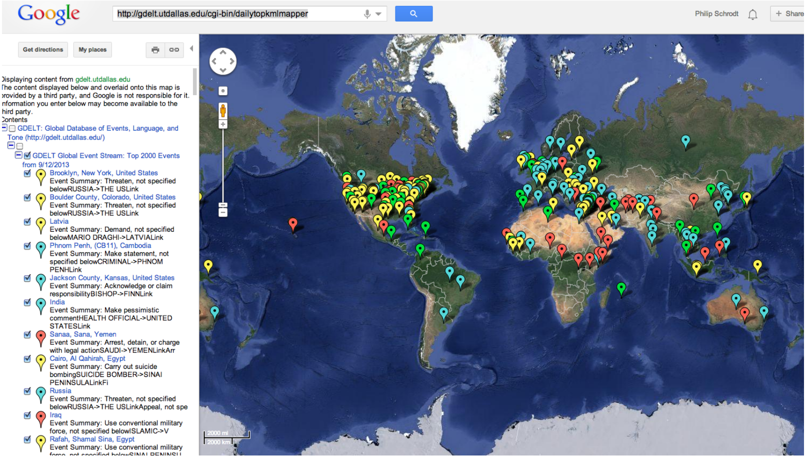
GDELT's goal is to help uncover previously-obscured spatial, temporal, and perceptual evolutionary trends through new forms of analysis of the vast textual repositories that capture global societal activity, from news and social media archives to knowledge repositories.

[Read More...]

**Rolf Fredheim's Visualization Tutorial**

Rolf Fredheim of Cambridge University put together a fantastic assortment of basic analyses and some sophisticated visualizations of GDELT, including a number of publication-friendly maps and graphs. He also shows how to create some fantastic network visualizations of actor interconnections. He includes all of the source code and descriptions of how he generated each image. [MORE FEATURES]

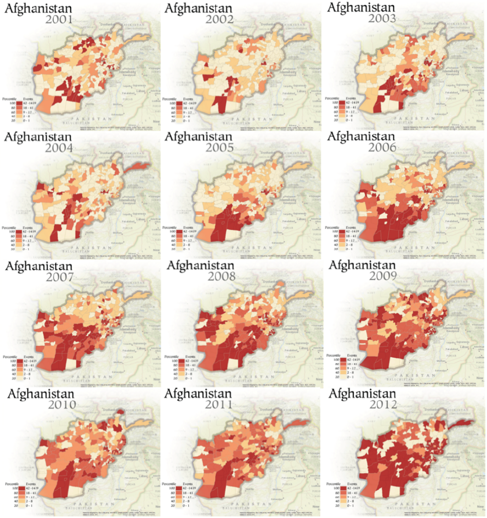
# http://gdelt.utdallas.edu



Displaying content from gdelt.utdallas.edu
The content displayed below and overlaid onto this map is provided by a third party, and Google is not responsible for it. Information you enter below may become available to the third party.
Contents

- □ GDELT: Global Database of Events, Language, and Tone (http://gdelt.utdallas.edu/)
  - □
    - ☑ GDELT Global Event Stream: Top 2000 Events from 9/12/2013
      - ☑ Brooklyn, New York, United States
        Event Summary: Threaten, not specified belowRUSSIA->THE USLink
      - ☑ Boulder County, Colorado, United States
        Event Summary: Threaten, not specified belowRUSSIA->THE USLink
      - ☑ Latvia
        Event Summary: Demand, not specified belowMARIO DRAGHI->LATVIALink
      - ☑ Phnom Penh, (CB11), Cambodia
        Event Summary: Make statement, not specified belowCRIMINAL->PHNOM PENHLink
      - ☑ Jackson County, Kansas, United States
        Event Summary: Acknowledge or claim responsibilityBISHOP->FINNLink
      - ☑ India
        Event Summary: Make pessimistic commentHEALTH OFFICIAL->UNITED STATESLink
      - ☑ Sanaa, Sana, Yemen
        Event Summary: Arrest, detain, or charge with legal actionSAUDI->YEMENLinkArr
      - ☑ Cairo, Al Qahirah, Egypt
        Event Summary: Carry out suicide bombingSUICIDE BOMBER->SINAI PENINSULALinkFi
      - ☑ Russia
        Event Summary: Threaten, not specified belowRUSSIA->THE USLinkAppeal, not spe
      - ☑ Iraq
        Event Summary: Use conventional military force, not specified belowISLAMIC->V
      - ☑ Rafah, Shamal Sina, Egypt
        Event Summary: Use conventional military force, not specified belowSINAI PENINSU

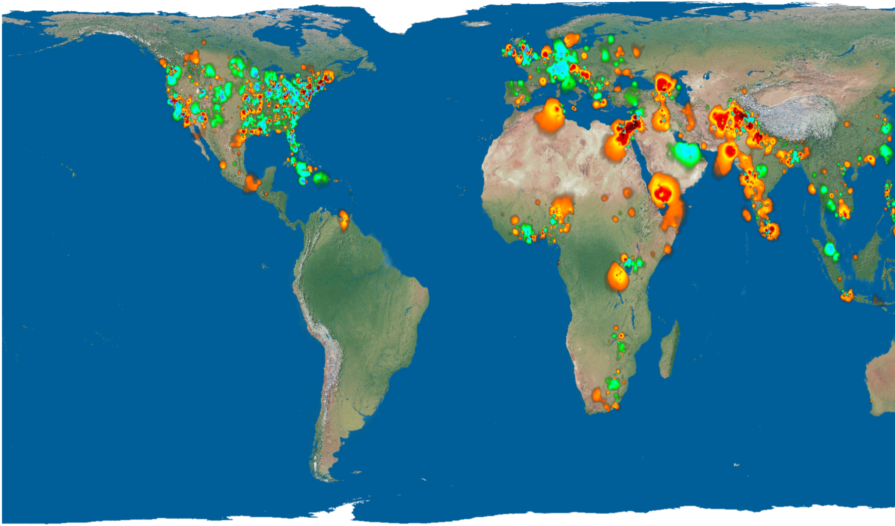
# Afghanistan, District-level Violence



[This is *not* Wikileaks data!]

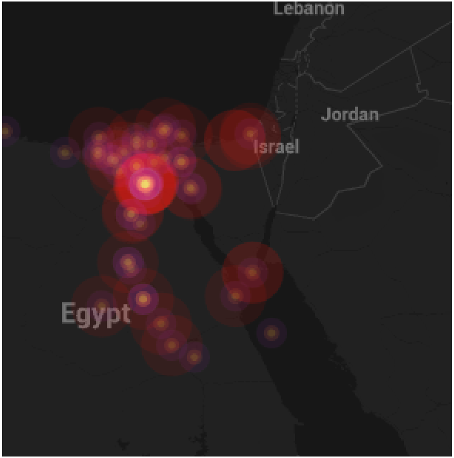
Source: Jay Yonamine and Joshua Stevens, Penn State

# Heat-map of Events, 29 Jan 2011



Source: Kalev Leetaru

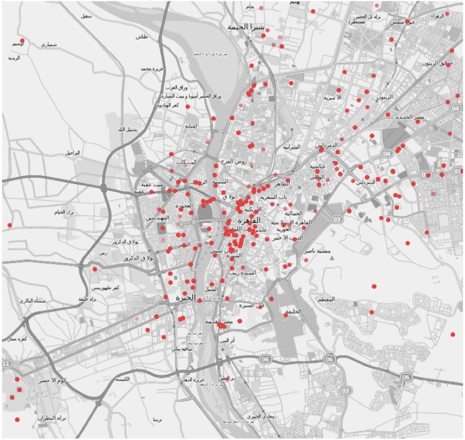
# Egypt protests: intensity



Source: John Beieler and Joshua Stevens, Penn State

# Cairo protests: location

GDELT Coverage of Events in Cairo, Jan 2010 to July 2013



Source: David Masad and Andrew Halterman of Caerus Analytics.

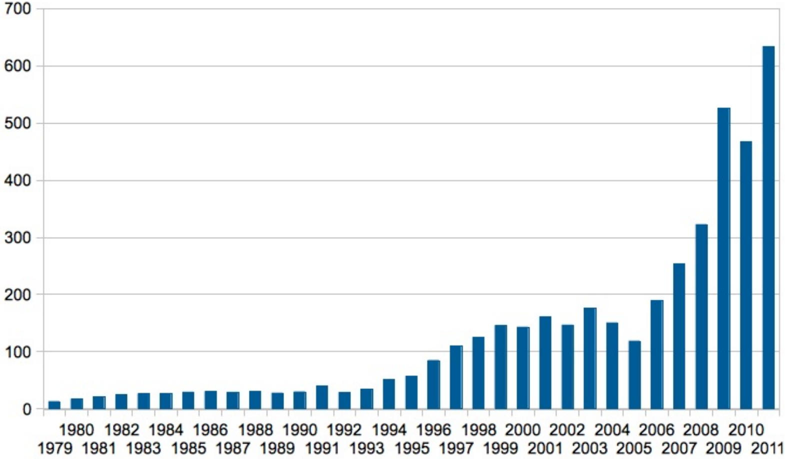
# Topics:

- Overview
- The Big Picture
- Visualizations
- Why GDELT is difficult to use
- GDELT 2.0
- Named Entity Recognition/Resolution
- Additional work to be done

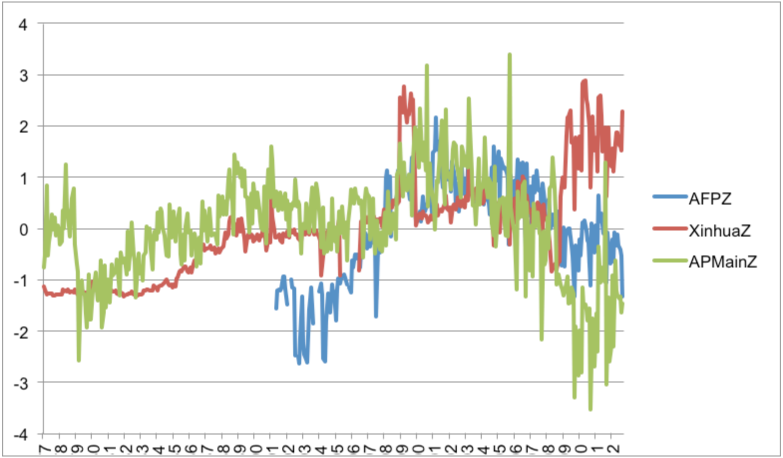
# GDELT is big, really, really big

- Downloading takes a while: please be patient with the servers
- It is too large to read entirely into *R* unless you really know what you are doing (though there is some debate about this)
- There is substantial redundancy in the variables: these can usually be reduced to the much smaller number which you are actually using
- You probably just want to use subsets anyway

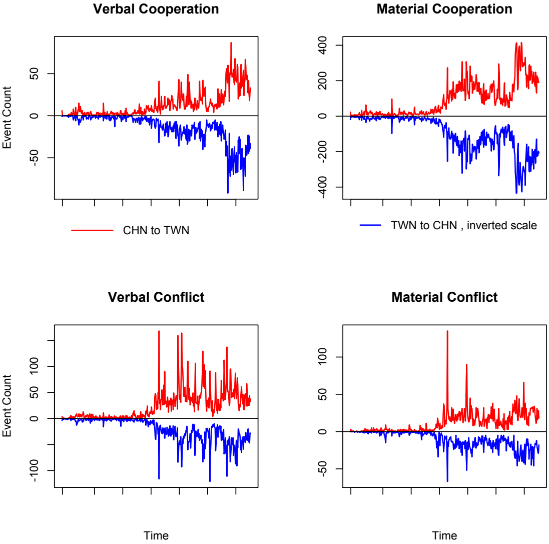
# Base rate increases exponentially over time



# Density of Data across Time, Core Sources, Z-scores

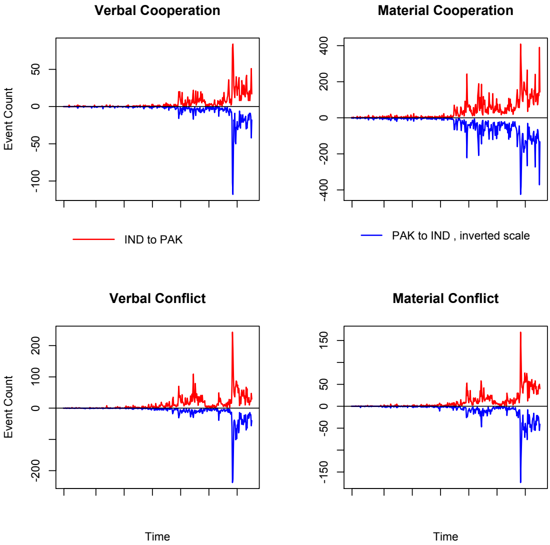


# Quad Counts, China → Taiwan



X-axis: 1979 to 2012

# Quad Counts, India → Pakistan



X-axis: 1979 to 2012

# Increase in coverage over time

Causes

- Sources being coded come in at different times
- Google News begins to come in around 2002
- Web-based sources have increased in general over the past ten years
- Data has the usual bias towards coverage of wealthy countries

Corrections

- Note that the *variance* increases as well as the mean
- Be very careful doing any sort of time-series analysis, particularly on the entire period
- Check the blog and other materials for various corrections people are using: this is still an open issue but normalizing by some function of the total number of events seems to work

# Very high number of false positives

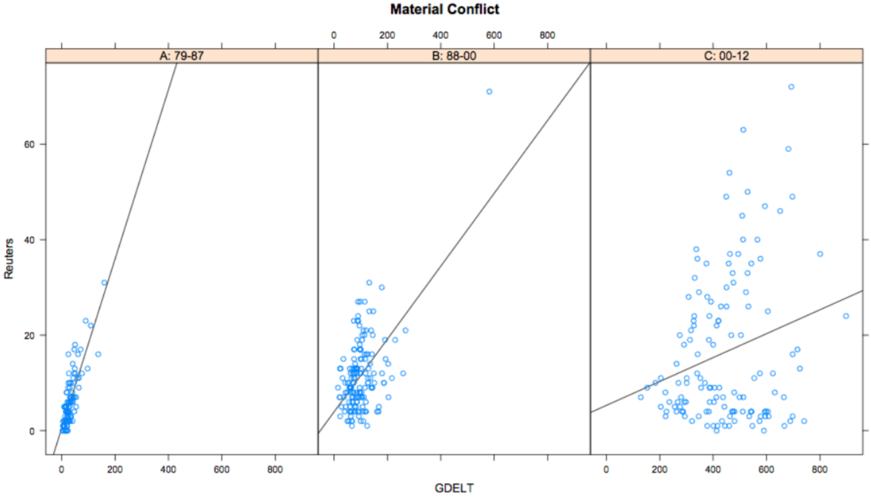
Causes

- The state-oriented CountryInfo.txt provides most of the dictionary
- Remaining dictionaries—including the verb phrases—are based on earlier KEDS/TABARI work, which was not global
- Full-story coding
- CAMEO dictionaries were not uniformly developed for all categories, particularly the 4-digit
- Kalev's *objective* was to extract as many events as possible, and in particular the system does not require full sources and targets

Corrections

- Use methods that are insensitive to false positives: there are many
- Never fail to warn potential users about false positives, even as they never fail to ignore this fact
- Filter

# Comparison with KEDS/Reuters: Israel → Palestine

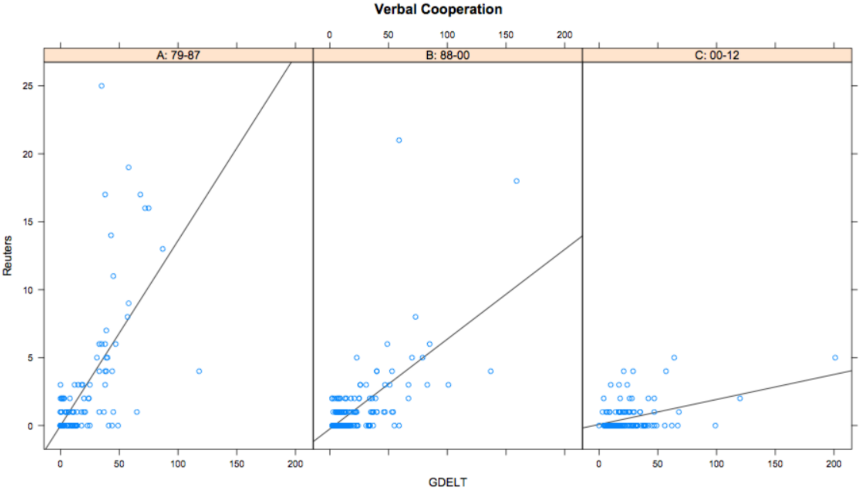


r = 0.812

r = 0.570

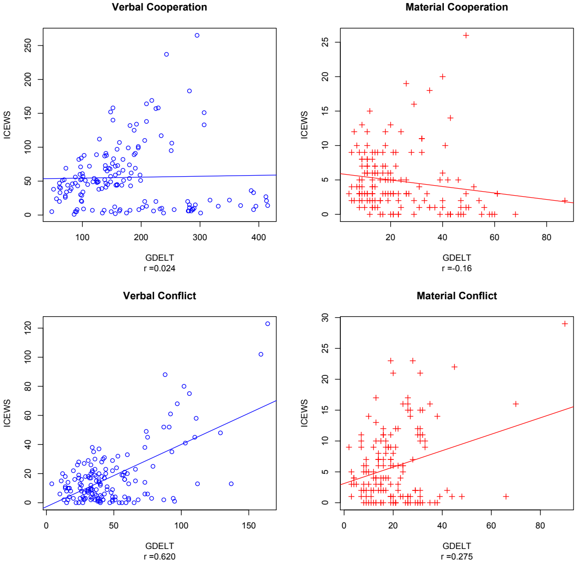
r = 0.234

# Comparison with KEDS/Reuters: Israel → Lebanon



r = 0.639 r = 0.631 r = 0.420

# Comparison with ICEWS Asia: China → Taiwan



Verbal Cooperation

r =0.024

Material Cooperation

r =-0.16

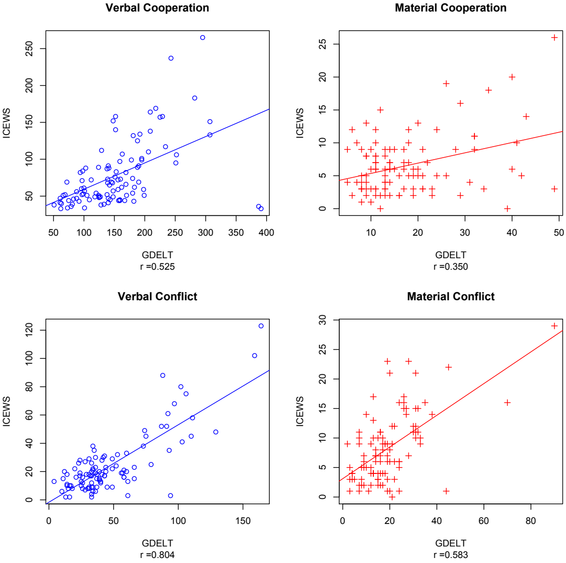
Verbal Conflict

r =0.620

Material Conflict

r =0.275

# Comparison with ICEWS Asia: China → Taiwan , high ICEWS months only



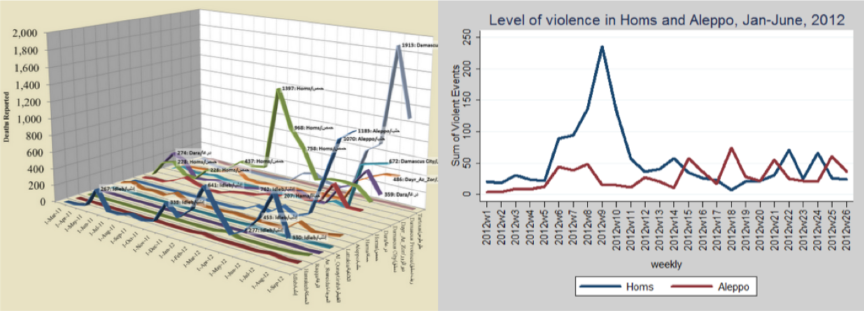
Verbal Cooperation

Material Cooperation

Verbal Conflict

Material Conflict

# Comparison with Syria Ushahidi data



Source: Jay Yonamine, Penn State

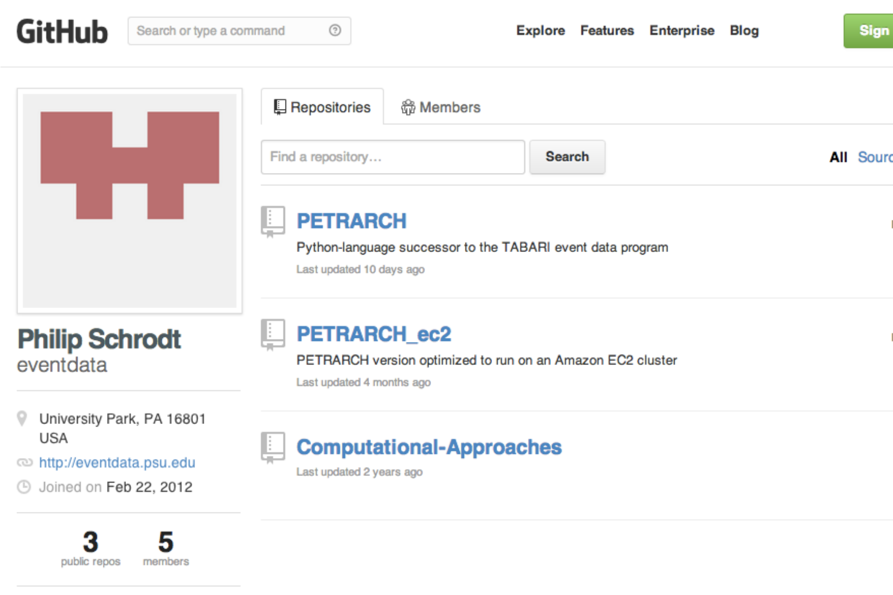
# Topics:

- Overview
- The Big Picture
- Visualizations
- Why GDELT is difficult to use
- GDELT 2.0
- Named Entity Recognition/Resolution
- Additional work to be done

# PETRARCH

New coding engine replacing TABARI

- Hosted on GitHub with multiple contributors
- Python rather than C/C++
    - Far larger—and younger—development community
    - Data structures are almost entirely textual and therefore more transparent



Philip Schrodt
eventdata

University Park, PA 16801
USA

http://eventdata.psu.edu

Joined on Feb 22, 2012

3 public repos

5 members

**PETRARCH**
Python-language successor to the TABARI event data program
Last updated 10 days ago

**PETRARCH_ec2**
PETRARCH version optimized to run on an Amazon EC2 cluster
Last updated 4 months ago

**Computational-Approaches**
Last updated 2 years ago

# Why Python?

- Open source (of course...tools want to be free...)
- Standardized across platforms and widely available/documented
- Automatic memory management (unlike C/C++)
- Generally more coherent than perl, particularly when dealing with large programs
- Text oriented rather than GUI oriented (unlike Java)
- Extensive libraries but these are optional (unlike Java):
- seems to be generating very substantial network effects
- C/C++ code can be easily integrated in high-performance applications
- Tcl can be used for GUI

# PETRARCH

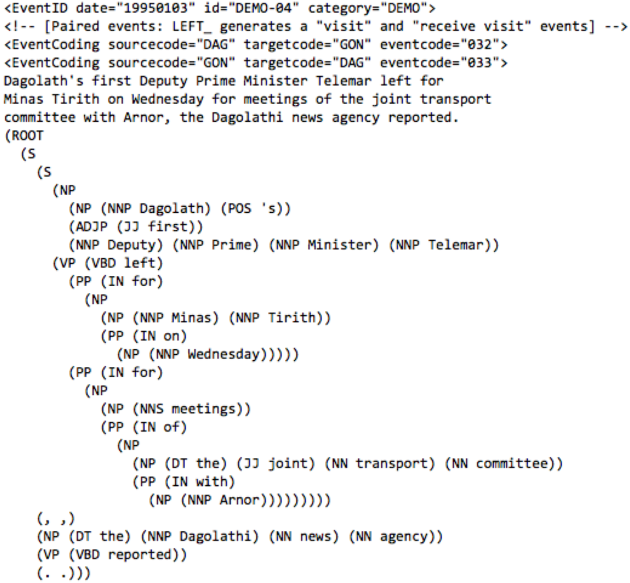
New coding engine replacing TABARI

- Hosted on GitHub with multiple contributors
- Python rather than C/C++
  - Far larger—and younger—development community
  - Data structures are almost entirely textual and therefore more transparent
  - Should work without modification across multiple operating systems, including Windows
- Well-documented "hooks" for adding alternative processing and utilities such as feature extractors
- Hosted on GitHub with multiple contributors
- Uses the output from the Stanford CoreNLP parser and coreferencing system
- Designed from the beginning for cloud processing

# Advantages of the CoreNLP parsing compared to TABARI shallow parsing

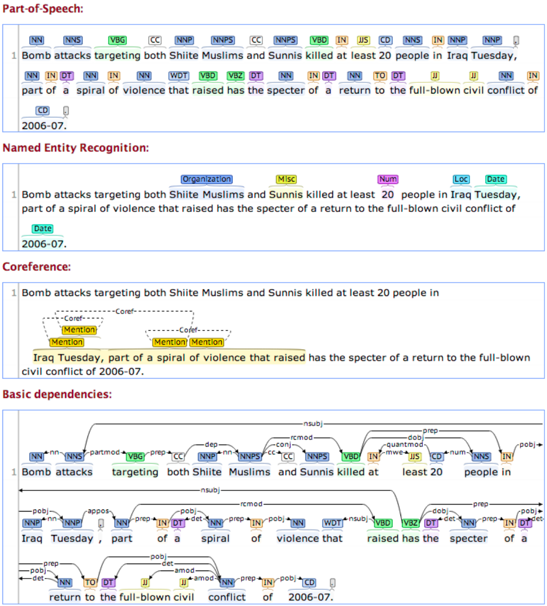
- Reduces incorrect identification of direct objects, which messes up source identification
- Provides noun/verb/adjective disambiguation:many words in English can be used in all three modes:
  - “A protest occurred on Sunday” [noun]
  - “Demonstrators protested” [verb]
  - “Marchers carried protest signs” [adjective]
- Identification of all named entities through noun phrases:
  - TABARI required actor to be in dictionaries.
  - PETRARCH will always pull these out whenever they occur in the source or target position;
  - The result unidentified cases can be separately processed with named-entity-resolution (NER) software
- More sophisticated co-referencing of pronouns and other references, particularly across sentences

# Stanford CoreNLP parse tree



```
<EventID date="19950103" id="DEMO-04" category="DEMO">
<!-- [Paired events: LEFT_ generates a "visit" and "receive visit" events] -->
<EventCoding sourcecode="DAG" targetcode="GON" eventcode="032">
<EventCoding sourcecode="GON" targetcode="DAG" eventcode="033">
Dagolath's first Deputy Prime Minister Telemar left for
Minas Tirith on Wednesday for meetings of the joint transport
committee with Arnor, the Dagolathi news agency reported.
(ROOT
  (S
    (S
      (NP
        (NP (NNP Dagolath) (POS 's))
        (ADJP (JJ first))
        (NNP Deputy) (NNP Prime) (NNP Minister) (NNP Telemar))
      (VP (VBD left)
        (PP (IN for)
          (NP
            (NP (NNP Minas) (NNP Tirith))
            (PP (IN on)
              (NP (NNP Wednesday)))))
        (PP (IN for)
          (NP
            (NP (NNS meetings))
            (PP (IN of)
              (NP
                (NP (DT the) (JJ joint) (NN transport) (NN committee))
                (PP (IN with)
                  (NP (NNP Arnor)))))))))
    (, ,)
    (NP (DT the) (NNP Dagolathi) (NN news) (NN agency))
    (VP (VBD reported))
    (. .)))
```

# Stanford CoreNLP word dependency and coreferences



**Part-of-Speech:**

1 Bomb attacks targeting both Shiite Muslims and Sunnis killed at least 20 people in Iraq Tuesday, part of a spiral of violence that raised has the specter of a return to the full-blown civil conflict of 2006-07.

NN NNS VBG CC NNP NNPS CC NNPS VBD IN JJS CD NNS IN NNP NNP .
NN IN DT NN IN NN WDT VBD VBZ DT NN IN DT NN TO DT JJ JJ NN IN
CD .

**Named Entity Recognition:**

1 Bomb attacks targeting both Shiite Muslims and Sunnis killed at least 20 people in Iraq Tuesday, part of a spiral of violence that raised has the specter of a return to the full-blown civil conflict of 2006-07.

**Coreference:**

1 Bomb attacks targeting both Shiite Muslims and Sunnis killed at least 20 people in Iraq Tuesday, part of a spiral of violence that raised has the specter of a return to the full-blown civil conflict of 2006-07.

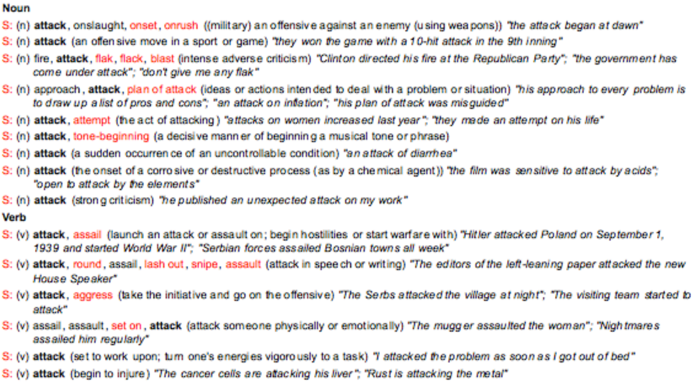
**Basic dependencies:**

1 Bomb attacks targeting both Shiite Muslims and Sunnis killed at least 20 people in Iraq Tuesday , part of a spiral of violence that raised has the specter of a return to the full-blown civil conflict of 2006-07.

# Problems PETRARCH/CoreNLP does not solve

- Word-sense disambiguation
    - "attack": physical or verbal?

# WordNet word senses: “attack”



**Noun**

- S: (n) **attack**, onslaught, onset, onrush ((military) an offensive against an enemy (using weapons)) *"the attack began at dawn"*
- S: (n) **attack** (an offensive move in a sport or game) *"they won the game with a 10-hit attack in the 9th inning"*
- S: (n) fire, **attack**, flak, flack, blast (intense adverse criticism) *"Clinton directed his fire at the Republican Party"; "the government has come under attack"; "don't give me any flak"*
- S: (n) approach, **attack**, plan of attack (ideas or actions intended to deal with a problem or situation) *"his approach to every problem is to draw up a list of pros and cons"; "an attack on inflation"; "his plan of attack was misguided"*
- S: (n) **attack**, attempt (the act of attacking) *"attacks on women increased last year"; "they made an attempt on his life"*
- S: (n) **attack**, tone-beginning (a decisive manner of beginning a musical tone or phrase)
- S: (n) **attack** (a sudden occurrence of an uncontrollable condition) *"an attack of diarrhea"*
- S: (n) **attack** (the onset of a corrosive or destructive process (as by a chemical agent)) *"the film was sensitive to attack by acids"; "open to attack by the elements"*
- S: (n) **attack** (strong criticism) *"he published an unexpected attack on my work"*

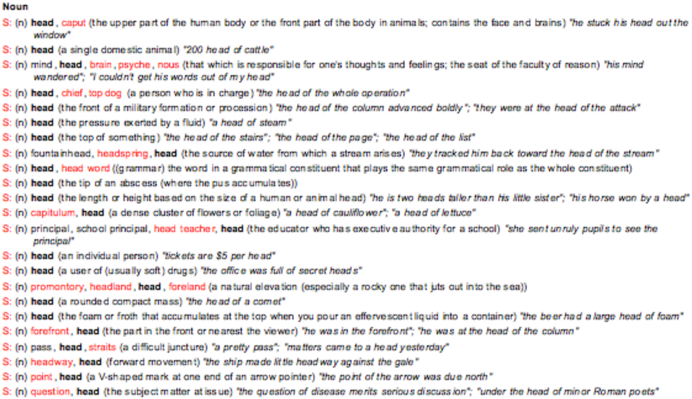
**Verb**

- S: (v) **attack**, assail (launch an attack or assault on; begin hostilities or start warfare with) *"Hitler attacked Poland on September 1, 1939 and started World War II"; "Serbian forces assailed Bosnian towns all week"*
- S: (v) **attack**, round, assail, lash out, snipe, assault (attack in speech or writing) *"The editors of the left-leaning paper attacked the new House Speaker"*
- S: (v) **attack**, aggress (take the initiative and go on the offensive) *"The Serbs attacked the village at night"; "The visiting team started to attack"*
- S: (v) assail, assault, set on, **attack** (attack someone physically or emotionally) *"The mugger assaulted the woman"; "Nightmares assailed him regularly"*
- S: (v) **attack** (set to work upon; turn one's energies vigorously to a task) *"I attacked the problem as soon as I got out of bed"*
- S: (v) **attack** (begin to injure) *"The cancer cells are attacking his liver"; "Rust is attacking the metal"*

# Problems PETRARCH/CoreNLP does not solve

- Word-sense disambiguation
  - "attack": physical or verbal?
  - “head” has about 65 different meanings in English, ranging from a leadership designation to a marine toilet.

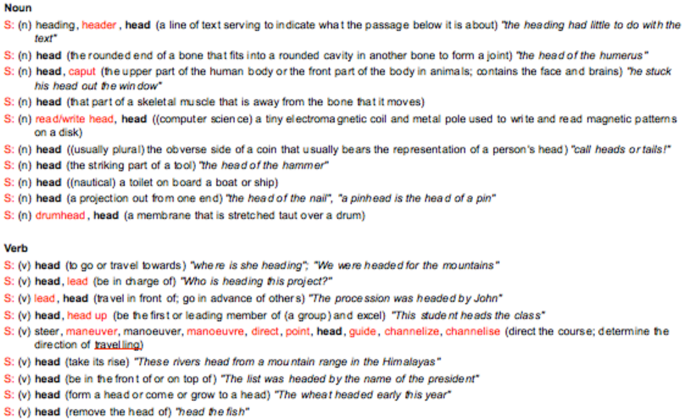
# WordNet word senses: "head"



**Noun**

- S: (n) **head**, caput (the upper part of the human body or the front part of the body in animals; contains the face and brains) *"he stuck his head out the window"*
- S: (n) **head** (a single domestic animal) *"200 head of cattle"*
- S: (n) mind, **head**, brain, psyche, nous (that which is responsible for one's thoughts and feelings; the seat of the faculty of reason) *"his mind wandered"; "I couldn't get his words out of my head"*
- S: (n) **head**, chief, top dog (a person who is in charge) *"the head of the whole operation"*
- S: (n) **head** (the front of a military formation or procession) *"the head of the column advanced boldly"; "they were at the head of the attack"*
- S: (n) **head** (the pressure exerted by a fluid) *"a head of steam"*
- S: (n) **head** (the top of something) *"the head of the stairs"; "the head of the page"; "the head of the list"*
- S: (n) fountainhead, headspring, **head** (the source of water from which a stream arises) *"they tracked him back toward the head of the stream"*
- S: (n) **head**, head word ((grammar) the word in a grammatical constituent that plays the same grammatical role as the whole constituent)
- S: (n) **head** (the tip of an abscess (where the pus accumulates))
- S: (n) **head** (the length or height based on the size of a human or animal head) *"he is two heads taller than his little sister"; "his horse won by a head"*
- S: (n) capitulum, **head** (a dense cluster of flowers or foliage) *"a head of cauliflower"; "a head of lettuce"*
- S: (n) principal, school principal, head teacher, **head** (the educator who has executive authority for a school) *"she sent unruly pupils to see the principal"*
- S: (n) **head** (an individual person) *"tickets are $5 per head"*
- S: (n) **head** (a user of (usually soft) drugs) *"the office was full of secret heads"*
- S: (n) promontory, headland, **head**, foreland (a natural elevation (especially a rocky one that juts out into the sea))
- S: (n) **head** (a rounded compact mass) *"the head of a comet"*
- S: (n) **head** (the foam or froth that accumulates at the top when you pour an effervescent liquid into a container) *"the beer had a large head of foam"*
- S: (n) forefront, **head** (the part in the front or nearest the viewer) *"he was in the forefront"; "he was at the head of the column"*
- S: (n) pass, **head**, straits (a difficult juncture) *"a pretty pass"; "matters came to a head yesterday"*
- S: (n) headway, **head** (forward movement) *"the ship made little headway against the gale"*
- S: (n) point, **head** (a V-shaped mark at one end of an arrow pointer) *"the point of the arrow was due north"*
- S: (n) question, **head** (the subject matter at issue) *"the question of disease merits serious discussion"; "under the head of minor Roman poets"*

# WordNet word senses: “head” continued



**Noun**

- S: (n) heading, header, **head** (a line of text serving to indicate what the passage below it is about) *"the heading had little to do with the text"*
- S: (n) **head** (the rounded end of a bone that fits into a rounded cavity in another bone to form a joint) *"the head of the humerus"*
- S: (n) **head**, caput (the upper part of the human body or the front part of the body in animals; contains the face and brains) *"he stuck his head out the window"*
- S: (n) **head** (that part of a skeletal muscle that is away from the bone that it moves)
- S: (n) read/write head, **head** ((computer science) a tiny electromagnetic coil and metal pole used to write and read magnetic patterns on a disk)
- S: (n) **head** ((usually plural) the obverse side of a coin that usually bears the representation of a person's head) *"call heads or tails!"*
- S: (n) **head** (the striking part of a tool) *"the head of the hammer"*
- S: (n) **head** ((nautical) a toilet on board a boat or ship)
- S: (n) **head** (a projection out from one end) *"the head of the nail", "a pinhead is the head of a pin"*
- S: (n) drumhead, **head** (a membrane that is stretched taut over a drum)

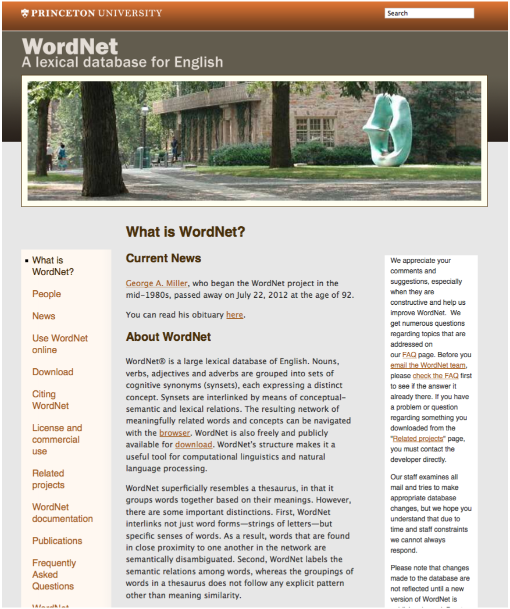
**Verb**

- S: (v) **head** (to go or travel towards) *"where is she heading"; "We were headed for the mountains"*
- S: (v) **head**, lead (be in charge of) *"Who is heading this project?"*
- S: (v) lead, **head** (travel in front of; go in advance of others) *"The procession was headed by John"*
- S: (v) **head**, head up (be the first or leading member of (a group) and excel) *"This student heads the class"*
- S: (v) steer, maneuver, manoeuver, manoeuvre, direct, point, **head**, guide, channelize, channelise (direct the course; determine the direction of travelling)
- S: (v) **head** (take its rise) *"These rivers head from a mountain range in the Himalayas"*
- S: (v) **head** (be in the front of or on top of) *"The list was headed by the name of the president"*
- S: (v) **head** (form a head or come or grow to a head) *"The wheat headed early this year"*
- S: (v) **head** (remove the head of) *"head the fish"*

# Problems PETRARCH/CoreNLP does not solve

- Word-sense disambiguation
  - "attack": physical or verbal?
  - “head” has about 65 different meanings in English, ranging from a leadership designation to a marine toilet.
- Detailed development (and extension) of the CAMEO categories and dictionaries
  - CAMEO was developed to study mediation, not as a general-purpose coding ontology
  - Converting the TABARI dictionaries from WEIS to CAMEO took about three academic-research-project-years
  - This is mundane, sloggy, labor intensive task on the same scale as a large human-coded data project
  - it is not the sort of big data sexy topic that funders are ready to throw gobs of open-source/open-access money at.

# *WordNet*-based dictionaries



PRINCETON UNIVERSITY

Search

**WordNet**

A lexical database for English

## What is WordNet?

- What is WordNet?
- People
- News
- Use WordNet online
- Download
- Citing WordNet
- License and commercial use
- Related projects
- WordNet documentation
- Publications
- Frequently Asked Questions

### Current News

George A. Miller, who began the WordNet project in the mid-1980s, passed away on July 22, 2012 at the age of 92.

You can read his obituary here.

### About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

We appreciate your comments and suggestions, especially when they are constructive and help us improve WordNet. We get numerous questions regarding topics that are addressed on our FAQ page. Before you email the WordNet team, please check the FAQ first to see if the answer it already there. If you have a problem or question regarding something you downloaded from the "Related projects" page, you must contact the developer directly.

Our staff examines all mail and tries to make appropriate database changes, but we hope you understand that due to time and staff constraints we cannot always respond.

Please note that changes made to the database are not reflected until a new version of WordNet is

Source: http://wordnet.princeton.edu/

# *WordNet*-based dictionaries

- Verb dictionaries have been completely reorganized around *WordNet* synonym sets (“synsets”)
- Verb-phrase patterns include synsets for common objects such as currency, weapons and quantities
- “Agents” dictionary for common nouns—for example “police”, “soldiers”, “president”—includes all *WordNet* synsets
- Dictionaries will be reformatted into a JSON data structure
- Additional dictionary enhancements carried forward from TABARI 0.8
  - regular noun and verb endings
  - all irregular verb forms
  - improved dictionaries for militarized non-state actors

# Named Entity Recognition/Resolution

- Locating and classifying phrases into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
- Examples: http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html.
- No general solution; approaches tend to be either
  - Rule and dictionary based, which requires manual development
  - Sequence-based machine-learning methods, specifically conditional random fields. These require an extensive set of marked-up examples
- *Name resolution* involves either
  - Differentiating two distinct entities which have the same name: “President Bush”
  - Combining multiple names of the same entity" “Obamacare” and “Affordable Care Act”
- Network models which associate a particular use of the name with other entities and/or time are frequently useful here.

# Named Entity Recognition/Resolution

Most research on NER systems has been structured as taking an unannotated block of text, such as this one:

*Jim bought 300 shares of Acme Corp. in 2006.*

And producing an annotated block of text, such as this one:

```
<ENAMEX TYPE="PERSON">Jim</ENAMEX>
  bought<NUMEX TYPE="QUANTITY">300</NUMEX>
   shares of
   <ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX
    in <TIMEX TYPE="DATE">2006</TIMEX>.
```

In this example, the annotations have been done using so-called ENAMEX tags that were developed for the Message Understanding Conference in the 1990s.

State-of-the-art NER systems for English produce near-human performance. For example, the best system entering MUC-7 scored 93.39% of F-measure while human annotators scored 97.60% and 96.95%.

Source: http://en.wikipedia.org/wiki/Named_entity_recognition

# Projected work by Kalev Leetaru

- “Tone”: this is the work to be done under Kalev’s Yahoo fellowship at Georgetown:
  - techniques for doing this go back to some of the earliest efforts in automated coding, notably Philip Stone’s *General Inquirer*
  - Possible problem: most news articles are intentionally neutral in tone
- Coding material in Google, Library of Congress and other sources back to 1800
  - Kalev is hoping this will give “cultural” information similar to Google NGram Viewer but with subject/object differentiation
  - Distinguishing fiction and non-fiction sources may be problematic
  - We will still have the “Mickey Mouse gap” from 1927-1980 or so

# Topics:

Overview

The Big Picture

Visualizations

Why GDELT is difficult to use

GDELT 2.0

Named Entity Recognition/Resolution

Additional work to be done

# Missing topics in CAMEO

- Routine democratic political processes
  - Elections
  - Legislative debate
- Human rights violations
- Criminal activity
  - Narcotics
  - Cyber crime
- Natural disasters (IDEA coding framework)
- Disease (IDEA coding framework)
- Financial crises and event-like discontinuities

More generally, CAMEO provides *too much* detail on mediation, which it was originally designed to code

# Specialized data sets

- Protest
  - Size
  - Topic[s]
  - Sponsor[s]
  - Response of authorities[s]
  - Location resolved below the city level
- Monitoring/situational awareness
  - Minimize the false positive rate
  - Quad-category only
  - Specialized categories only, e.g. events possibly related to climate change

Major issue: how can we integrate dictionaries produced at multiple sites to maximize the total coverage?

# Increasing the speed and efficiency of dictionary development

- NER systems for near-real-time updating of actors and open collaboration on maintenance of major actor dictionaries
- Automated identification of new verb phrases: we've never tried this
- Cloud-sourcing elements of dictionary development and validation
- Establishing a "ground truth" validation set covering all of the CAMEO categories
- Standardization of religion, ethnic groups and militarized non state actors

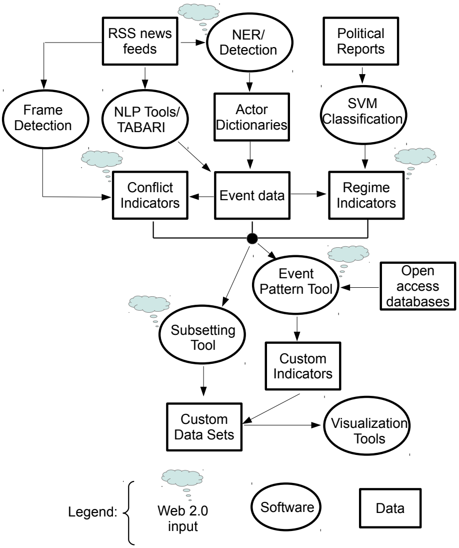
# Expanding local coverage

- Locating sources which are either open access or have non-predatory licensing arrangements
  - Event-to-source “drill-down” is a very high priority
  - Sources need to be shared across projects even if they are not open
  - al-Jazeera?
  - “Wikinews”?
- Non-English sources, probably through Google Translate or a comparable system
- Location-specific dictionaries for actors and events
- Utilize NGO sources to the extent that this is ethical and secure

# Sharing computational requirements and tool development

- GDELT 2.0 is highly computationally intensive compared to GDELT 1.0/TABARI
  - Hosting and subsetting
  - Parsing
  - NER
  - Translation
- Shared tool development, on the other hand, seems to be going very well
- We anticipate developing a textbook-like instructional tool/site and possible some MOOC-like video materials

# MADCOW


RSS news feeds
NER/ Detection
Political Reports
Frame Detection
NLP Tools/ TABARI
Actor Dictionaries
SVM Classification
Conflict Indicators
Event data
Regime Indicators
Event Pattern Tool
Open access databases
Subsetting Tool
Custom Indicators
Custom Data Sets
Visualization Tools
Legend:
Web 2.0 input
Software
Data

# Questions?

Contact:

schrodt735@gmail.com

Data: http://gdelt.utdallas.edu

Blog: http://gdeltblog.wordpress.com

TABARI/CAMEO: http://eventdata.parusanalytics.com

PETRARCH and other tools: https://github.com/eventdata