

# A Practical Guide to Generating Event Data

Philip A. Schrod  
Pennsylvania State University  
[schrod@psu.edu](mailto:schrod@psu.edu)  
<http://eventdata.psu.edu>

# Outline

- Existing data sets
- Unix and programming
- Downloading and formatting
- Named entity recognition
- Coding with TABARI
- Actor and event ontologies
- Aggregation
- High-volume coding

# Existing data sets

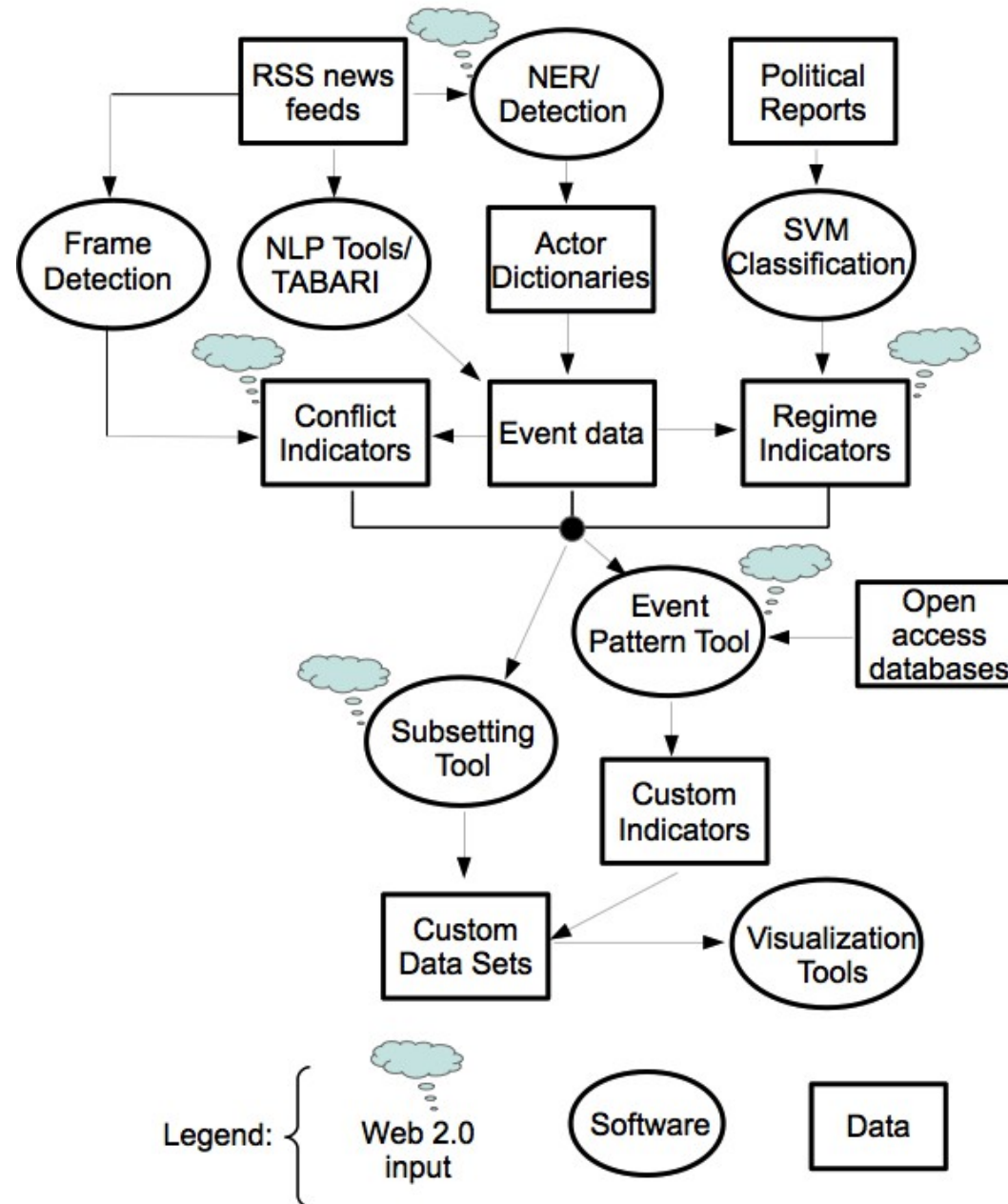
**TABLE 1** Selected Event Data Projects

Dataset	Focus	Geographical	Years	Geo-located?	Reference
ACLED	Conflict	Primarily Africa	1997–2010	Yes	Raleigh et al. 2010
EDACS	Violence	Failed states	1990–2009	Yes	Chojnacki et al. 2012
GTD	Terrorism	Global	1970–2010	Yes	START 2012
ICEWS	General	Asia; global	1998–2010	No	O'Brien 2010
KEDS	General	Primarily Middle East	1979–2011	No	Schrodt and Gerner 2010
KOSVED	One-sided violence	Selected states	varies by case	No	Schneider et al. 2012
MID3 Incidents	Conflict	Global	1993–2001	Yes; MID- LOC	COW 2007; Braithwaite 2010
NIRI	Violence	Northern Ireland	1968–1998	Yes	Sullivan et al. 2012
PITFWAD	One-sided violence	Global	1995–2012	Yes	PITF 2011
SCAD	Protest	Africa	1990–2010	Yes	Salehyan et al. 2012
SIGACTS	Violence	Afghanistan, Iraq	2004–2010	Yes	HSRP 2010; Linke et al. 2012
SPEED	General	Global	1946–2010	City	Nardulli 2011
UCDP-GED	Organized violence	Africa	1989–2010	Yes	Melander and Sundberg 2011
UCDP/PRIO ACD	Conflict	Global	1946–2011	Conflict site	Themnér and Wallensteen 2011; Hallberg 2012
Urban Violence	Urban disorder	Africa, Asia	1960–2009	Yes	Urdal and Hoelscher 2012
VRA	General	Global	1990–2004	No	King and Lowe 2006
WITS	Terrorism	Global	2004–2010	City	NCTC 2011
WARICC	Water-related conflict	Mediterranean, Middle East, Sahel	1997–2009	Yes	Bernauer et al. 2012

# Event Model: Core Innovation

- Once calibrated, real-time event forecasting models can be run ***entirely*** without human intervention
  - Web-based news feeds provide a rich multi-source flow of political information in real time
  - Statistical models can be run and tested automatically, and are 100% transparent
- In other words, for the first time in human history—quite literally—we have a system that can provide real-time measures of political activity without any human intermediaries

# Integrated open, real time data generation



# Event Data in the [original] DARPA period

- Nation-state orientation; most analysis dealt with Cold War major power relations
- Global coverage
- Human coding
- Source texts were major Western newspapers
- Statistical models were relatively simple

# Contemporary Event Data

- Substate and nonstate actors; most analysis deals with protracted conflicts
- Single-conflict and regional coverage
- Automated coding
- Source texts are from wire services (Reuters, AFP, BBC)
- Statistical models are very complex

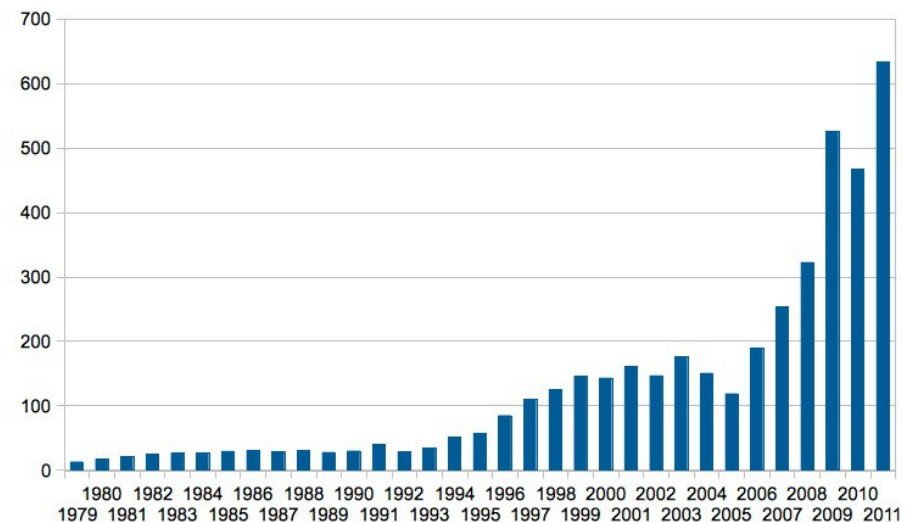
# ICEWS Phase I Event Data

- 30-gigabytes of text from Lexis-Nexis
- 25 sources
- 8-million stories
- 26-million sentences
  - Only first four sentences coded in each story
- 3-million events
- Generally two orders of magnitude greater than any prior event coding effort



# Forthcoming Event Data

- Global coding, Jan-1979 to Jun-2012
- 180-million events based on open news sources
- CAMEO event, actor and sub-state agent coding
  - 15,000 verb phrase dictionary
  - 40,000+ political actors and agents
- Geolocated to city level
- Planned quarterly updates, backfit to 1900 or possibly 1800



## What we know about event data in 2012 that *confirms* what we suspected in 1975

- Media fatigue is a major factor in event reporting
- WEIS and its derivatives—IDEA and CAMEO—contains most of the major categories required to code political interactions
- Human coding has about 25% error rate in long-term projects even when coders are initially trained to 90%+ accuracy. Multi-year, multi-institution projects may have *substantially* lower accuracy
- It is impossible for human coders to keep up with coding in real time
- Comments and meetings are about 30% to 50% of most event data
- Violent events are reported disproportionately

# What we know about event data in 2012 that we *didn't know* in 1975

- Machine coding to a level of accuracy comparable to (or better than) multi-institution human coding teams is straightforward
- Vast quantities of news reports are available in machine-readable form and can be downloaded automatically and for free using RSS feeds
- Some of the original WEIS categories cannot be consistently differentiated
- Scales and detailed coding categories add relatively little information; event reports alone explain about 50% to 75% of the variance
  - (but journal editors keep telling authors to remove this statistical finding from articles accepted for publication)

# What we know about event data in 2012 that we *didn't know* in 1975, continued

- News sources vary dramatically in their coverage; these effects differ by region and time. However, news service reports provide substantially greater coverage than individual newspapers
  - There appears to have been an exponential increase in the number of available stories starting about 2005.
  - Following the model of the study of pre-modern systems, we can apply what we've learned from conflicts where we have good data to conflicts where the data is not as good.
- Regionally specific data sets provide better coverage than global data data sets. It is difficult to maintain consistent coverage across the entire international system
  - This assessment may change with the newer global data sets

# Pre-requisites

- “With great power comes great responsibility”

# Pre-requisites

- “With great power comes great responsibility”
- Sourced variously to Luke 12:48, Voltaire, Gandalf, and Albus Dumbledore but in fact...

# Pre-requisites

- “With great power comes great responsibility”
- Sourced variously to Luke 12:48, Voltaire, Gandalf, and Albus Dumbledore but in fact...
- Stan Lee, Spiderman I

# Why do we have to learn all this technical crap?!?!

Subtext: I just want to study politics!!!

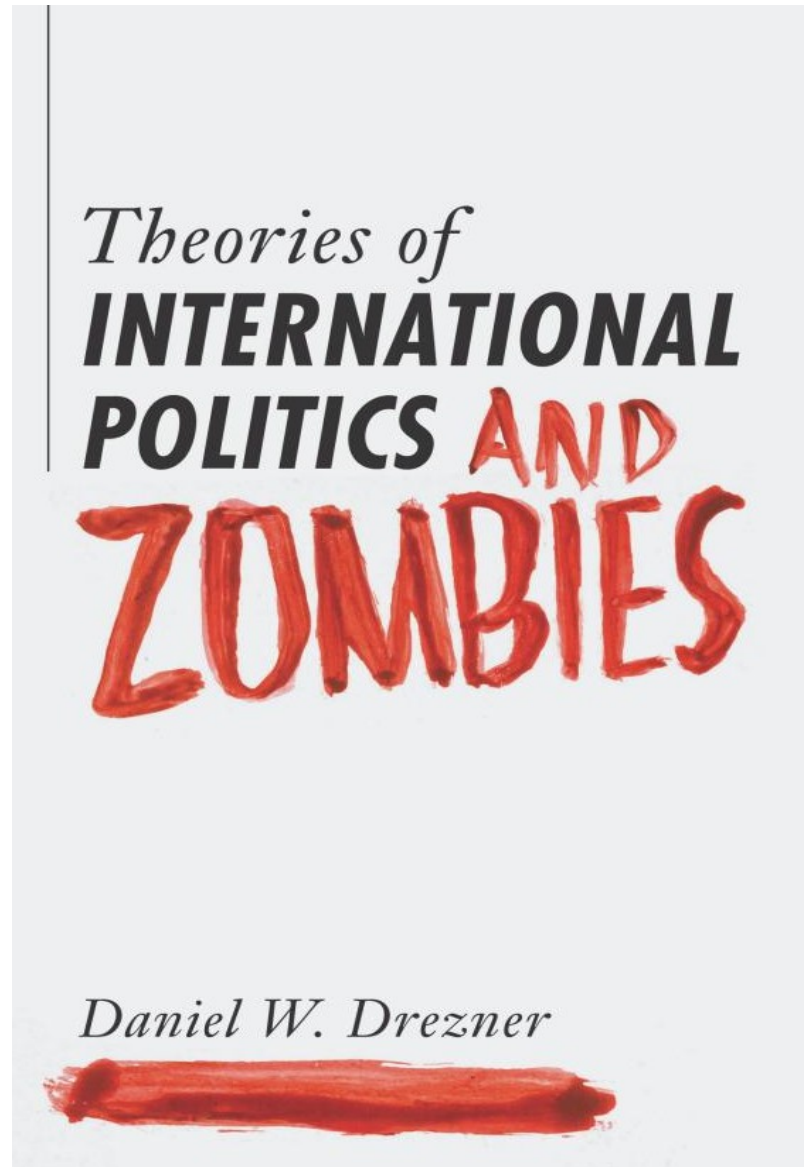
- Hey, dude, we're "scientists"...
- Could be worse...you could be a medical student...
  - Or a medieval historian
- The past fifty years have, in fact, marked a transition from where political science "research tools" consisted of a comfortable chair and a snifter of brandy to where we can make effective use of highly complex machines
  - Burt Monroe's legislative debate data sets at Penn State are second in size to the Sloan Sky Survey of the entire universe
- None of this is likely to change any time soon



This?...



Or this...



# Computing Power

Control Data Corporation 3600  
(ca.1965)  
32 K (48-bit) RAM memory  
1 processor  
~1-million operations per second  
Output: line printer



# Computing Power

Control Data Corporation 3600  
(ca.1965)  
32 K (48-bit) RAM memory  
1 processor  
~1-million operations per second  
Output: line printer



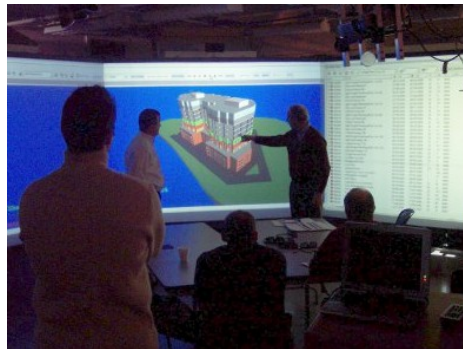
Penn State High Performance Computing Facility  
15 cluster computers  
100 to 2000 2.66 Ghz processors in each cluster  
~50 Gb RAM accessible to each processor  
130 Tb disk space  
4 interactive visualization rooms

# Computing Power

Control Data Corporation 3600  
(ca.1965)  
32 K (48-bit) RAM memory  
1 processor  
~1-million operations per second  
Output: line printer



Penn State High Performance Computing Facility  
15 cluster computers  
100 to 2000 2.66 Ghz processors in each cluster  
~50 Gb RAM accessible to each processor  
130 Tb disk space  
4 interactive visualization rooms



Motorola Razzr  
16 Gb RAM memory  
Dual processor  
~500-million operations per sec  
540 x 860 color display



# Open Source Software



WORDPRESS.COM





# Statistics Packages :: Cars

Stata



R



SAS



SPSS





R:



- Open source



- Open source
- Widely used in all approaches to statistical analysis *and* pattern recognition
- CRAN library provides almost immediate access to new methods
  - events (event data aggregation)
  - tm (text miner)
- Robust scripting capabilities; easily interfaces with C/C++ when needed
- Skill set is widely available



Cutting Edge: 1912



Cutting Edge: 2012

# NLP Components :: Smart Phones



# NLP... it's all components



# Evolution of Statistical Training in Political Science

- 1985: Objective is to get grad students to take advanced courses in econometrics
  - Hint: deadends— SEM, co-integration models

# Evolution of Statistical Training in Political Science

- 1985: Objective is to get grad students to take advanced courses in econometrics
  - Hint: deadends— SEM, co-integration models
- 2010: Objective is to develop advanced techniques that will be adopted in other disciplines
  - Imai, Sekhon, Fowler, Gill, King
  - Hint: Medical schools are rumored to pay better than the liberal arts

# The Transition We Need in Programming

- Old model:
  - “We’ll just hire a programmer because that will be more efficient than doing it ourselves”



# The Transition We Need in Programming

- Old model:
  - “We’ll just hire a programmer because that will be more efficient than doing it ourselves”
- Reality
  - Computer science departments and ExxonMobil can’t find enough programmers either
  - You take a serious efficiency hit in trying to explain what you want done
  - You may take a serious efficiency hit in not doing the task in the best way—NLP (and statistics) are specialized subfields
  - Programmers frequently are trained to focus on GUIs, (e.g. Java) which usually just get in the way in research computing

# Why people don't want to be programmers

- Programming is a craft, not a science
  - “Between the mathematics that make [the computer] theoretically possible and the electronics that makes it practically feasible lies the programming that makes it intellectually, economically and socially useful. Unlike the extremes, the middle remains a craft, technical rather than technological, mathematical only in appearance.”  
Michael Sean Mahoney, Histories of Computing (Harvard University Press)
  - Practice, practice, practice
- Programmer efficiency varies by a factor of 10 to 20, which can be very demoralizing
- Popular perception of programmers



# Why people don't want to be programmers

- Programming is a craft, not a science
  - “Between the mathematics that make [the computer] theoretically possible and the electronics that makes it practically feasible lies the programming that makes it intellectually, economically and socially useful. Unlike the extremes, the middle remains a craft, technical rather than technological, mathematical only in appearance.”  
Michael Sean Mahoney, Histories of Computing (Harvard University Press)
  - Practice, practice, practice
- Programmer efficiency varies by a factor of 10 to 20, which can be very demoralizing
- Popular perception of programmers



# We need to utilize...

- Some formal training in core algorithms and data structures
  - Not just AP Java
- Rapid development scripting languages: perl and Python
- Lingua franca (and GUI): Java
- High performance: C/C++
- Regular expressions are to text analysis what calculus was to modeling physical systems

# 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)
- C/C++ can be easily integrated by high-performance applications
- Tcl can be used for GUI

# Why use Unix?

- Stable and compact
  - About 20 commands will do almost everything you need; they haven't changed in 30 years
- Core of all major operating systems except Windows
  - Linux, OS-X, Android
  - Most functions are identical across OS-X and Linux
- Standardized set of compilers, so identical code can run on multiple systems.
  - “make” command will compile code on any machine
- Used in most cluster computers
- Research software is more likely to be written for Unix
- Command line is more efficient than mouse/menus in advanced applications

# TABARI Basics: #1 Problem people have running TABARI

Are your files in Unix format, not Windows or the old Macintosh format?

- If you saved them in Excel, they are not
- If you processed them on a Windows system, they are not
- There are a variety of ways to solve this—I use BBEdit—but you have to address it

(Or send me the modified code for TABARI to address it automatically)

# TABARI Basics: #2 Problem people have running TABARI

See problem #1



TABARI Basics: #3 Problem  
people have running TABARI

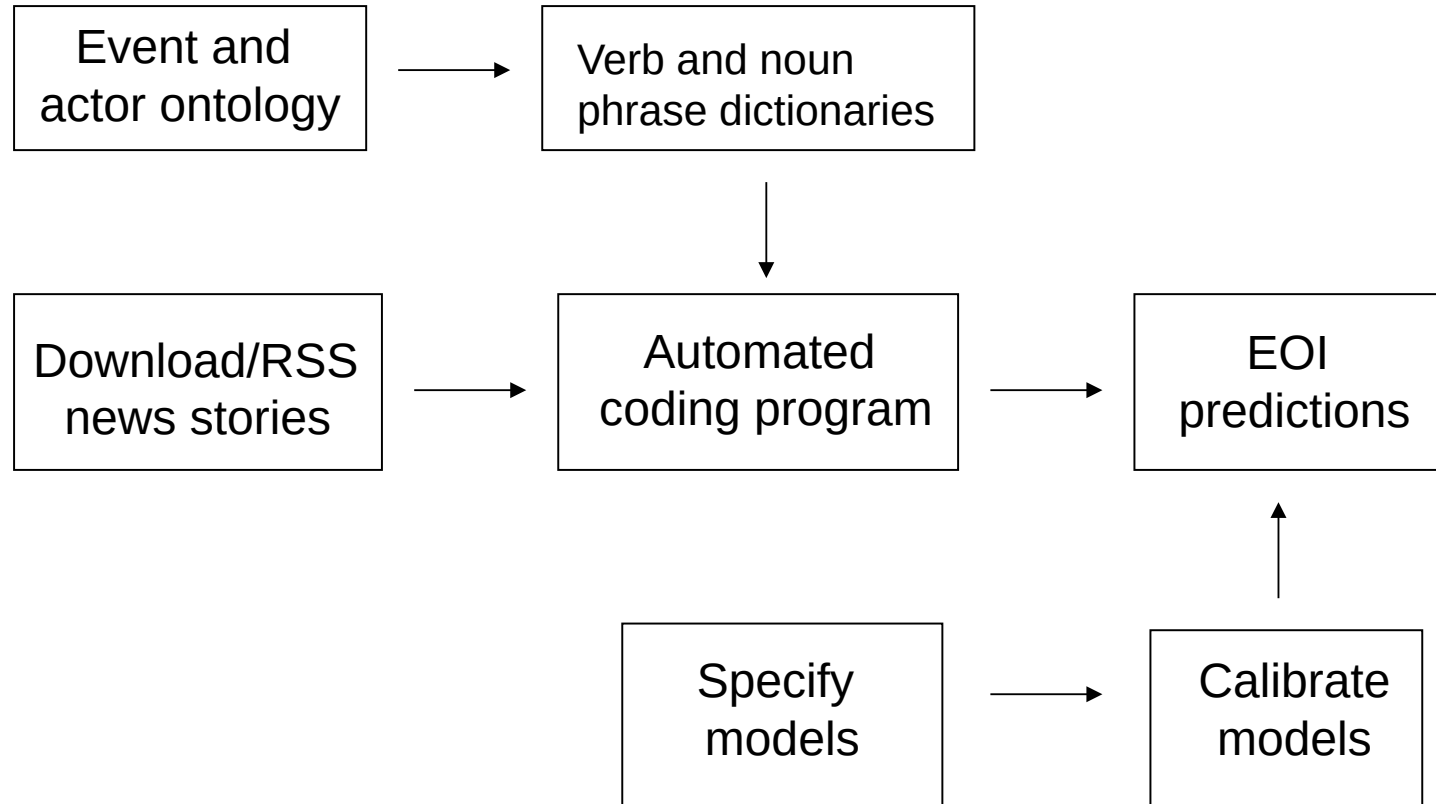
**R T F M**



## Basic challenges that make text analysis hard

---

# Event Data Generation Process



# Definition of an “event”: KEDS Project 1992

An event is an interaction which can be described in a natural language sentence which has as its subject and direct or indirect object an element of a set of actors, and as the verb an element of a set of actions, all of which are transitive verbs, and which can be associated with a specific point in time.

# Text in TABARI format

090201 AFPN-0001-01

Olympic swimmer Michael Phelps admitted Sunday that he had engaged in "regrettable" behavior and "demonstrated bad judgment" after a photo was published that appeared to show him smoking cannabis.

090201 AFPN-0002-01

Israeli aircraft bombed Gaza on Sunday hours after its leaders vowed to deal a blow to Hamas in response to new rocket fire two weeks following the end of a bloody war in the Palestinian enclave.

090201 AFPN-0003-01

The world economic crisis spells the death of globalization and action is needed to protect the poor, said organizers of the World Social Forum as it wrapped up in Brazil on Sunday.

090201 AFPN-0004-01

Israeli aircraft on Sunday bombed an empty police station in central Gaza and tunnels along the Hamas-ruled enclave's southern border with Egypt, witnesses said.

090201 AFPN-0005-01

Palestinian president Mahmud Abbas lashed out at his Hamas rivals on Sunday, as officials from Palestinian groups gathered in Cairo amid hopes of bolstering a ceasefire in the Gaza Strip.

# Text in TABARI format

090201 AFPN-0008-01

Israel vowed to strike back at Hamas on Sunday after renewed rocket fire from the Islamists' Gaza stronghold two weeks after the end of a bloody war in the battered Palestinian territory.

090201 AFPN-0009-01

US Middle East envoy George Mitchell met Saudi Arabia's King Abdullah on Sunday on the last leg of a Middle East tour aimed at reviving the Israeli-Palestinian peace process.

090201 AFPN-0010-01

Israeli Defence Minister Ehud Barak and Foreign Minister Tzipi Livni on Sunday voiced conflicting views on an emerging Egyptian-brokered ceasefire with Hamas in the Gaza Strip.

090201 AFPN-0011-01

Pro-Palestinian groups meeting at the World Social Forum that ended Sunday launched a call for a worldwide boycott of Israel and a day of action on March 30.

090201 AFPN-0012-01

The political supremo of Palestinian Islamist movement Hamas, Khaled Meshaal, on Sunday ruled out any "permanent ceasefire" until Israel ends its crippling blockade of the Gaza Strip.

# What it actually looks like

```
META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=utf-8">
<html>
<head>
<title>Full Article</title>
<style type="text/css">
<!--body { font-size: 100%;}-->
</style>

</head>
<body bgcolor="#FFFFFF" text="#000000">
<table width="100%" border="0" cellspacing="0" cellpadding="5" align="center">
  <tr> <td class="k9"><a href="http://www.factiva.com"></img></a></td>
  <td>
    <div align="right"></img></div>
  </td></tr>

</table> <table width="100%" border="0" cellspacing="0" cellpadding="5"
align="center">
<tr>
  <td valign="top" colspan="2" align="right" bgcolor="#FFFFFF" class="source"> <hr> </td>
</tr>
</table>
</td></tr>
</table>
```

# What it actually looks like

```
<table width="100%" border="0" cellspacing="0" cellpadding="5" align="center">
  <tr><td bgcolor="#d4e899" HEIGHT="25"><a name ="1"><div align="right"><a href
="#2"><font face="Verdana, Arial, Helvetica, sans-serif" size="1">
Next</font></a></div></td></tr> </table>
```

```
<table width="100%" border="0" cellspacing="0" cellpadding="5" align="center">
```

```
<tr><td><font face="Verdana, Arial, Helvetica, sans-serif" size="2"><br>
```

```
<b>Egypt police kill African migrant at Israel border</b><br>
```

```
LBA0000020091201e5c1000ww<br>
```

```
260 Words<br>
```

```
01 December 2009<br>
```

```
08:20 GMT<br>
```

```
Reuters News<br>
```

```
English<br>
```

```
(c) 2009 Reuters Limited </font></td></tr>
```

```
<tr><td width="90%"> <font face="Verdana, Arial, Helvetica, sans-serif" size="2">
```

```
<P> CAIRO, Dec 1 (Reuters) - Egyptian police shot and killed an African migrant on Tuesday as
he tried to slip across the Sinai peninsula desert border to Israel, a security source said.</P>
```

```
<P>Egyptian police have stepped up efforts in recent months to control the frontier with Israel
following an increase in human trafficking through Egypt. At least 17 migrants have been killed
at the border since May, the latest one two weeks ago.</P>
```

```
</font></td></tr>
```

```
<tr><td width="90%">
```



# What it actually looks like

<P> The Sinai border is on one of the main routes for African migrants and refugees, almost all unarmed, seeking work or asylum in Israel. Egyptian police say the smugglers who ferry migrants to the border region sometimes fire on security forces.</P>

<P> The security source said police did not know the dead man's nationality, but he appeared to be in his early twenties. Eritreans are the largest group of people trying to cross into Israel from Egypt, but Ethiopians and Sudanese also make the trek.</P>

<P> Analysts and aid workers say the flow of migrants from the Horn of Africa through Egypt to Israel has increased in recent months as it has become more difficult to travel on other northward routes, such as via Libya to Europe.

(Reporting by Rasha Kamal; writing by Yasmine Saleh; editing by Tim Pearce)

((yasmine.saleh@reuters.com ; +20 2 2578 3290; Reuters

Messaging: yasmine.saleh.reuters.com@reuters.net))</P>

<P></P>

</font></td></tr> <tr><td><font face="Verdana, Arial, Helvetica, sans-serif" size="2">

BE-EGYPT-BORDER/AFRICAN|LANGEN|AFA|CSA|LBY|RWSA|RWS|REULB|GNS|G|RBN|  
MD|AFN|RNP|DNP|PGE|SXNA<br>

</font></td></tr> </table>

# Text Sources

# International news sources

- Reuters
- Agence France Presse (AFP)
- BBC (various news feeds)
- AP and UPI
- Xinhua (excellent Africa coverage)
- New York Times
- Washington Post

# International news sources

- This is very much a moving target due to the changing dynamics of news gathering
- Coverage differs dramatically by region: “follow the money”
- Generally only available 1990 (or so) to the present; NYT Historical is an exception
- Most of these have web pages where you can get current content, but past content is limited

# Lexis-Nexis and Factiva

- No one has figured out how to automate the downloads of these
  - Instead, get your search string as precise as possible to reduce unnecessary downloads, then slog away...
- The LN search engine is very unpredictable: it is not designed for this sort of thing. Money (as in “DARPA”) was insufficient to solve the problem
  - Note that this contradicts the advice immediately above about refining the search string
- Factiva is far more reliable but somewhat more awkward to use
  - Money could solve that problem, but more money than you probably have.

# European Media Monitor

- Project of the EU's Joint Research Center
- Monitors over 4000 sites from 1600 key news portals world-wide plus 20 commercial news feeds and, for some applications, also specialist sites.
- Retrieves over 40000 reports per day in 43 languages.
- Classifies all news according to hundreds of subjects and countries.
- Access on the web, via email and by RSS.
- Runs 24 hours per day, 7 days a week.

Source: <http://emm.jrc.it/overview.html>

# Google News

- 4500 English-language sources
- Appears to have a facility for duplicate detection
- It's Google...

# What about new social media?



Mainstream media



# What about new social media?



Mainstream media



Internet and  
new social media

# New social media

- The good
  - Widely available to elites
  - More or less uncensored in open many societies
  - Should provide early information on changing sentiment prior to observing actual collective action
- The bad
  - No filters and mostly politically irrelevant:  
“Wanna getta pizza? ;)”
  - Easily manipulated by anyone—business, government, NGOs—who wants to go to the trouble of doing so
- The ugly
  - No standardization of content
- Utility in prediction
  - Multiple studies show this seems to work in the 6 to 48 hour range

# New Social Media

Police tell high school students to disguise their identity on FaceBook

+

Students choose the first country they recognize in an alphabetical list....Afghanistan

---

OMG! Jihadis are checking out the junior prom!

# New social media in text analysis



Mainstream media



Internet and  
new social media

## Number of stories found with “Palestinian killed” NEXIS search string

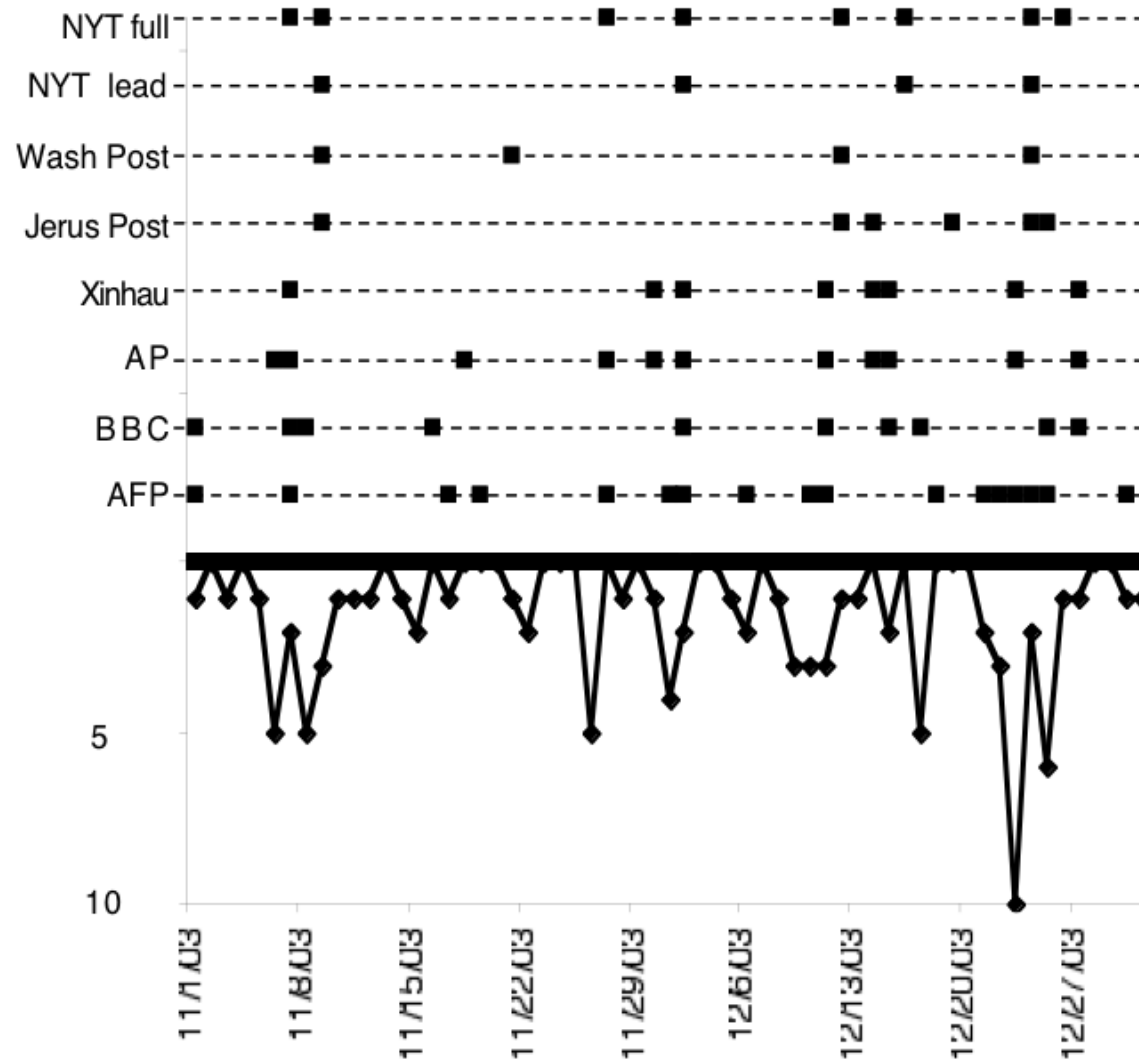
### **Newspaper**

• <i>Los Angeles Times</i>	3
• <i>New York Times</i>	4
• <i>Washington Post</i>	4
• <i>Jerusalem Post</i>	6
• <i>New York Times</i> , full text	8

### **Wire Service**

•Xinhau	8
•BBC (Factiva)	10
•Associated Press	11
•Agence France Presse	18

## Comparison of newspaper and wire service coverage of Palestinian deaths, Nov-Dec 2003



# Duplicate detection: sources

- Use of newswires
  - This is the reason for newswires
  - Detecting exact duplicates is easy
- Updating and corrections of previous stories
- News summaries
- Chronologies
- Multiple independent sources (Reuters, AFP, BBC)
- Except for exact duplicates, this is a difficult problem
  - “Near duplicate detection
  - One-a-Day filter
- Duplication probably amplifies the signal you want: a story is more likely to be repeated if it is important

# How many sources are needed?

- Probably 80% can be obtained from the major international sources
- Local sources do not necessarily give you better coverage of important political events
- Censorship (official or voluntary) is an issue
- International coverage follows the money
- There is probably a lot of regional variation on this
- Same for non-English sources: possibility of using machine translation



# Source and consistency of text

How much work will be involved in getting the text into a form you can use?

- ASCII/UniCode text, for example news reports
- HTML, but web formats change frequently
  - Python has modules for removing HTML tags: don't try doing this just with regular expressions
- PDF
  - Try the tm module in R
- Scanned/OCR text
- Proprietary word processing formats (Word)
- New media sources such as blogs and tweets

# Style of language

- News reports and official documents are usually formal, syntactically-correct English
- Quotations and letters are a mix of formal and informal
- Open-ended responses range from formal to very fragmentary
- New media sources are often very informal and abbreviated
- Variants of English and changes in usage over time (e.g. slang, memes)
- Languages other than English

# Intellectual property

- Copyright law is generally open to “fair use” in research and education. However, institutional contracts with data providers are more limited
  - Information does not necessarily want to be free
- Human subjects considerations—and therefore IRB review—apply to identifiable data
- A lot of the legal issues, particularly involving content on the web, are still *very* open
  - You probably do not want to be a test case
  - Just because someone claims IP rights doesn't mean they actually have those right
  - You still probably do not want to be a test case
  - Public institutions have considerable protection from sovereign immunity, though many have been wimps in asserting this.

# And of course, costs

- Is the information source already formatted?
  - Spinn3R, Thomas
  - Web pages vary dramatically in ease of downloading
- How much data do you actually need?
  - Text data sets are frequency much, much larger than typical political science data such as surveys and national indicators
  - Will just a sample be sufficient?
  - Are you coding more information than you will actually use?
- How much time will it take to code each document?
  - Who's going to train and supervise the manual coders?
  - How much can be fully automated?
  - How good is good enough?

TABARI

# 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

# Additional automated coding systems

- VRACoder (Bond, VRA, Cambridge, MA)
  - IDEA coding system
  - Deep parsing
  - Proprietary, Windows environment
- PERICLES (Shellman, Strategy Analysis Enterprises)
  - CAMEO coding system, dictionaries
  - Multi-field capabilities
  - Open-source, C#/.NET environment
- JABARI-NLP (Lockheed Martin)
  - TABARI structure with open-source parsing
  - Some geolocation capabilities; extensive substate actor coding

# MIT Robotics 101: First exercise

“Design a robot to wash dishes”



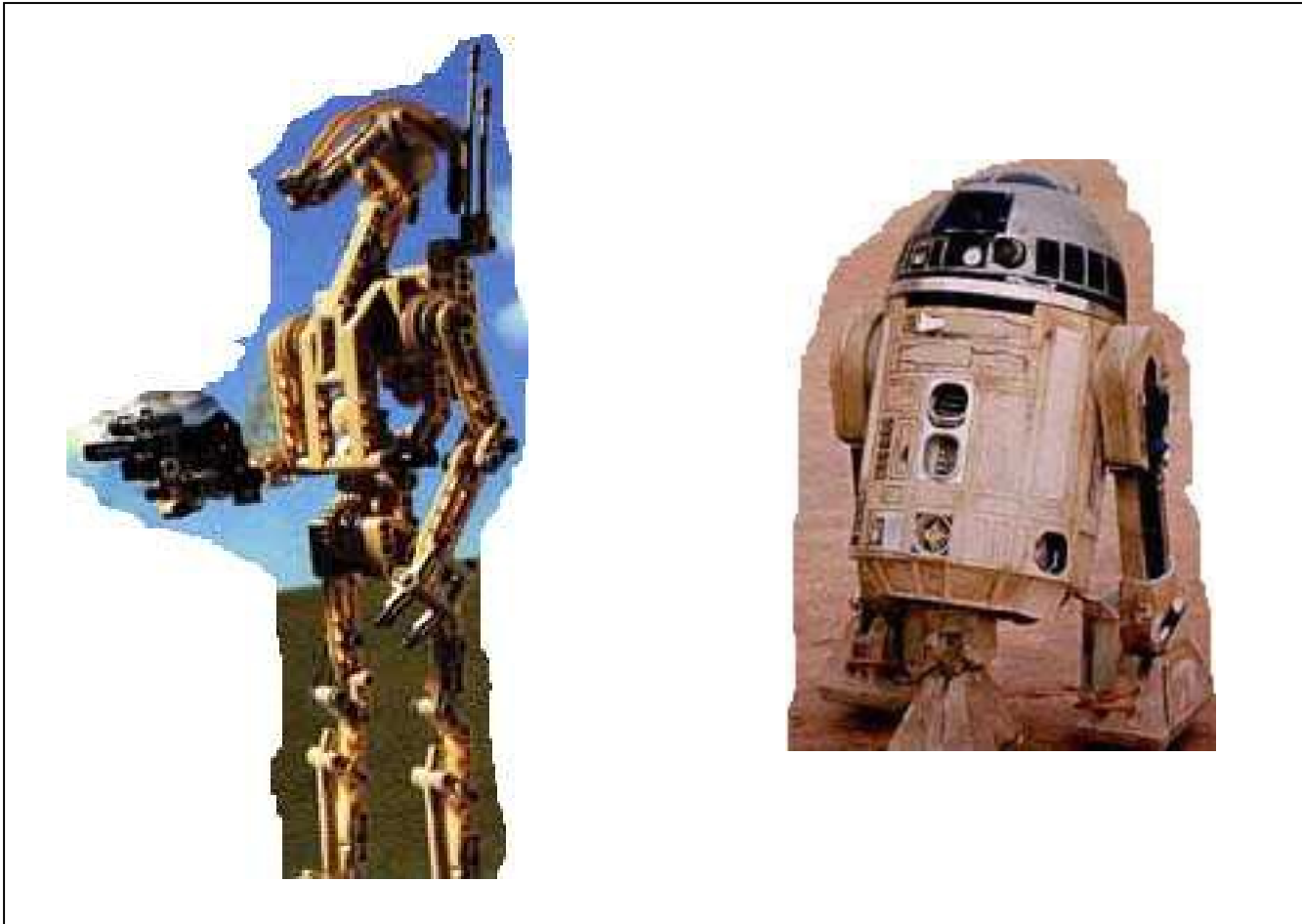
# Robotics 101: First Exercise

## Most frequent answer



# Robotics 101: First Exercise

## Slightly more appropriate answers



# Robotics 101: First Exercise

## Correct answer



# Event Coding systems

- WEIS
  - Charles McClelland, Rodney Tomlinson, DARPA
- COPDAB
  - Edward Azar
- PANDA
  - Doug Bond
- IDEA
  - Doug Bond, Craig Jenkins and Charles Taylor
- CAMEO
  - Deborah Gerner and Philip Schrod

# Categorization of Political Interactions

- Distinct English-language verb phrases:

5,000 to 10,000

(MUC, KEDS, PANDA projects)

- Micro-level categories

50 to 150

(WEIS, BCOW, IDEA, CAMEO)

- Macro-level categories

10 to 20

(WEIS, COPDAB, IPB, World Handbook)

# CAMEO: Event Coding

- Combines ambiguous categories in WEIS (promise/agree, grant/reward, warn/threaten)
- Eliminates WEIS subcategories for which no examples could be found
- Substantially expands coding for acts of violence
- Coding categories can be expanded to three levels
  - Originally designed for coding mediation but subsequently generalized for coding actions of militarized non-state actors
- Complete coding manual with examples of all event categories
- Implemented with a 15,000 verb phrase dictionary

# CAMEO: Actor Coding

- Systematic hierarchical scheme for coding sub-state and non-state actors
- Typical full actor code has three levels
  - State
  - Role
  - Identity
- Example: Hamas is coded PSEREBHMS
  - PSE: ISO-3166-alpha-3 code for the West Bank and Gaza
  - REB: Militarized opposition group
  - HMS: individual code for Hamas
- Additional rules standardize the coding of IGOs, NGOs, government leaders and so forth

# CAMEO Actor Coding: Block 1

- ISO-3166-1-alpha 3 country codes
- Religious/ethnicity codes
  - HURIDOCS religion codes
  - Maoz/Henderson religion codes?
  - ??? ethnicity codes
- Generic International/Transnational Actor Codes

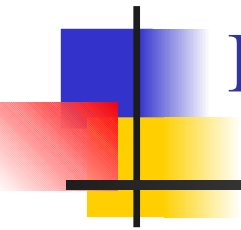


# CAMEO Actor Coding: Block 2

- Generic Domestic Actor/Role codes  
Modified from PANDA/IDEA codes
- Sub-state Region codes  
Possibility of using various UN and  
ISO geographical sources that code  
economic data
- Religious/ethnicity codes
- Special International/Transnational Actor Codes
- Country codes for NGOs and MNCs

# CAMEO Actor Coding: Block 3

- Generic Domestic Actor/Role codes
- Religious/ethnicity codes
- Branches of international organizations
- Special Actor Codes



# Human and Automated Coding

---

# Reliability in content analysis

- **Stability**—the ability of a coder to consistently assign the same code to a given text;
- **Reproducibility**—intercoder reliability;
- **Accuracy**—the ability of a group of coders to conform to a standard.

Source: Weber (1990:17)

# Advantages of automated coding

- Fast and inexpensive
- Transparent: coding rules are explicit in the dictionaries
- Reproducible: a coding system can be consistently maintained over a period of time without the “coding drift” caused by changing teams of coders.
- Coding dictionaries can be shared between institutions
- The coding of individual reports is not affected by the biases of individual coders. Dictionaries, however, can be so affected.
- It is possible to create rules for difficult technical and cultural vocabulary that is otherwise difficult to learn

# Disadvantages of automated coding

- Automated thematic coding has problems with disambiguation
- Automated syntactic coding makes errors on complex sentences by incorrectly identifying the object of the sentence.
- Requires a properly formatted, machine-readable source of text, therefore older paper and microfilm sources are difficult to code.
- Development of new coding dictionaries is time-consuming—KEDS/PANDA initial dictionary development required 2-labor-years. (Modification of existing dictionaries, however, requires far less effort)

# Human and machine coding tradeoffs

- Machine coding uses only information that is explicit in the text; human coders are likely to use implicit knowledge of the situation.
- Machine coding is not affected by boredom and fatigue
- Human coders can more effectively interpret idiomatic and metaphorical text, provided they are familiar with the context
- Human coders can more effectively deal with complex subordinate phrases and other unexpected grammatical constructions

# Summary

## **Advantage to human coding**

- Small data sets
- Data coded only one time at a single site
- Existing dictionaries cannot be modified
- Complex sentence structure
- Metaphorical, idiomatic, or time-dependent text
- Money available to fund coders and supervisors

## **Advantage to machine coding**

- Large data sets
- Data coded over a period of time or across projects
- Existing dictionaries can be modified
- Simple sentence structures
- Literal, present-tense text
- Money is limited



# Implications for automated coding

- Old objective:

Machine coding should attempt to duplicate human coders

- (which, in fact, can be done: Schrodtt and Gerner 1994, Bond et al 1997, Thomas 2001, King and Lowe 2003)
- Also human coding accuracy is probably *much* worse than the 80% that is usually claimed

- Alternative objective:

Optimize coding systems and models to use information that can be coded most reliably by machine

# Word frequency in English

<u>% of usage</u>	<u># of words</u>
40%	50
60%	2,300
85%	8,000
99%	16,000

- Total words in American English: about 600,000
- Total words in technical English (all fields):  
about 3-million

# Functional Words

Very short words such as

- Articles: a an the
- Interrogatives: who what when where why how
- Prepositions: to from at in above below
- Auxillary verbs: have has was were been
- Markers: by in at to (French de, German du, Arabic fi)
- Pronouns: I you he she him her his hers

In English, the specificity of a word is *generally* proportional to its length. These short will typically be in the stop word list, though a few longer words (e.g. “though” and “although”) also will be stop words

Marker words have multiple uses: *Random House College Dictionary* lists 29 meanings for “by,” 31 for “in,” 25 for “to,” and 15 for “for.”

# It gets harder: Disambiguation (“Bat”)

- Noun

- wooden (or aluminum) cylinder used in the game of baseball
- small flying mammal

- Verb

- act of batting (“at bat”)
- blinking (“bat an eye”)

- Idiomatic uses

- “go to bat for”: defending or interceding;
- “right off the bat”: immediately;
- “bats in the belfry”: commentary on an individual’s cognitive ability

- Foreign phrases

- “bat mitzvah”: a girl’s coming-of-age ceremony (Hebrew).

# WordNet word senses: “bat”

## Noun

- S: (n) **bat**, **chiropteran** (nocturnal mouselike mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate)
- S: (n) **bat**, **at-bat** ((baseball) a turn trying to get a hit) *"he was at bat when it happened"; "he got four hits in four at-bats"*
- S: (n) **squash racket**, **squash racquet**, **bat** (a small racket with a long handle used for playing squash)
- S: (n) **cricket bat**, **bat** (the club used in playing cricket) *"a cricket bat has a narrow handle and a broad flat end for hitting"*
- S: (n) **bat** (a club used for hitting a ball in various games)

## Verb

- S: (v) **bat** (strike with, or as if with a baseball bat) *"bat the ball"*
- S: (v) **bat**, **flutter** (wink briefly) *"bat one's eyelids"*
- S: (v) **bat** (have a turn at bat) *"Jones bats first, followed by Martinez"*
- S: (v) **bat** (use a bat) *"Who's batting?"*
- S: (v) **cream**, **bat**, clobber, **drub**, **thrash**, **lick** (beat thoroughly and conclusively in a competition or fight) *"We licked the other team on Sunday!"*

# Disambiguation, cont.

- Any of these uses might be encountered in an English-language text. Multiple uses might be found in a single sentence:

“The umpire didn’t bat an eye as Sarah lowered her bat to watch the bat flying around the pitcher.”

# Disambiguation, cont.

- Words can also change from verbs to nouns without modification.

Consider

- I plan to drive to the store, then wash the car.
- When John returned from the car wash, he parked his car in the drive.

- In summary:

“Verbing weirds language.”

Bill Watterson, *Calvin and Hobbes*

# 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"*



# 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"*

# Memes, idioms, metaphors and slang

- Political text frequently uses distinct idiomatic phrases
  - “Right to life”, “right to choice”
- Memes can have a high frequency for brief periods of time
  - “lipstick on a pig”
  - “top kill”, “junk shot”, “Deep Horizon”
- Military metaphors are common in political (and sports) rhetoric
  - “Tea Party insurgency”, “battleground state”
- OMG! WTF! Like IMHO slang expressions are common, and rapidly changing, in new media (lol...)

# Pre-processing

- We see distinct words, sentences, punctuation and make distinctions between entities and actions.
- A computer sees none of this... and more
- It must be instructed to distinguish what we think is important – “tokenizing”

## Further processing

- Next step is to get machine to pay attention to the tokens that are relevant to our goals – and to ignore those that are not.
- Explicitly delineate relevant tokens – keywords
- Remove generally irrelevant stuff
  - Stop words
  - Stemming

Note: In event data coding, this is usually done implicitly in the dictionaries rather than as an explicit processing step

# Stop words and Stemming

## Stop words

- a about above across after  
again against all almost  
alone along.... take taken  
than that the their them then  
....

## Stemming

- Suffix stripping
  - “ing”
- N-grams
  - “post office”
- Lemmatization
  - “meeting”
- It's complicated!

# Lessons from ICEWS and GDELT: dictionary calibration

- Existing verb-phrase dictionary required relatively little modification
- Actor dictionaries would benefit from re-structuring into general and country specific dictionaries
  - General: international actors, both state and non-state
  - General: discard phrases
  - General: WordNet-based agent dictionaries
  - KEDS, ICEWS: Specific national and regional actors
  - GDELT: CountryInfo.txt

# TABARI Project file

Coding file for events for ICEWS Nov09 coding, 10 November 2009

<dochead> Text source: leads only. Source: Agence France Press and Reuters

<doctail> For further information on the TABARI automated coding program and CAMEO

<doctail> coding system, consult <http://web.ku.edu/keds>

<verbsfile> CAMEO.091023.master.verbs

<actorsfile> nouns\_adj\_null.091023.txt

<actorsfile> Countries.091023.actors

<actorsfile> Internatnl.090916.rev.actors

<agentfile> ICEWS.091023sorted.agents

<issuesfile> GTDS.issues

<optionsfile> CAMEO.09b5.options

<textfile> LN.Texts0911.97.1

<textfile> LN.Texts0911.97.2

<textfile> LN.Texts0911.98.1

<textfile> LN.Texts0911.98.2

<textfile> LN.Texts0911.99.1

<textfile> LN.Texts0911.99.2

<textfile> LN.Texts0911.99.3

<eventfile> ICEWS.Nov09.pt1.evt

<coder> PAS

<session 1> <coder PAS> <date Wed 11 Nov 2009> <start 11:41 CST> <end 12:10 CST>

<records 6933308> <actorchanges 0> <verbchanges 0> <last 0>

<session 2> <coder PAS> <date Wed 11 Nov 2009> <start 12:17 CST> <end 12:43 CST>

<records 6933308> <actorchanges 0> <verbchanges 0> <last 0>



# TABARI Verbs Dictionary

ABDUCT [181];shan 7/24/93

- + \* AT\_ GUNPOINT FROM CAR \$ [181];sls 28 Mar 2008
- + HOSTAGE \* IN\_ \$ FREED [0841];sls 28 Mar 2008
- \* \_AND\_KILLED BY\_ \$ [1823];sks 18 Mar 2008
- \* AFTER ATTACK [181];ab 31 Dec 2005
- + HOSTAGE \* IN\_ \$ [181];sks 18 Mar 2008
- \* OF\_ + IN\_ \$ [181];JON 5/26/95
- + \* BY\_ \$ [181];BERT 12/31/97
- + IN\_ \$ { WAS | WERE } \* [181];sls 21 Sep 2007

ABIDE [---];shan 2/20/93

- WOULD NOT \* [120]
- FAIL \* [120];JON 5/8/95

ABSOLVE [---];ab 31 Dec 2005

- + WAS\_ \* IN\_ \$ [080];HUX 12/28/97

ABSTAIN [---];jw 10/25/91

- \* FROM VIOLENCE [087];ab 10 Nov 2005

ACCELERAT [---];sls 24 Sep 2007

- \* DEPLOYMENT TO\_ + [154];SLS 28 Nov 2007
- \* BUILDUP FORCE [154];sls 24 Sep 2007

ACCEPT [---];jw 11/14/91

- MUST\_ \* OF\_ + GOVERNMENT [102];SPK 30 Jan 2008
- \* LOAN FROM\_ + [071];SPK 07 Mar 2008
- WILL DECIDE WHETHER TO\_ \* [014]
- \* REQUEST JOIN MISSION [031];ab 10 Nov 2005
- \* ELECTION\_ DEADLINE [083];ai 05 May 2008
- \* \_RESPONSIBILITY [015];jap 25 Jun 2003
- SAID WOULD NEVER \* [120]
- CANNOT \* PROPOSAL [120];ab 31 Dec 2005
- \* PLAN TO\_ RESOLVE [037];sls 01 Mar 2008
- NOT \* PEACEKEEP [1245];BNL 06 May 2003
- LEANED TOWARD \* [---];jap 25 Jun 2003
- \* JURISDICTION [063];ab 10 Nov 2005

# TABARI Actors Dictionary

ABADAN [IRN] ;JON 5/24/95  
ABASSI\_MADANI [DZAREBFIS] ; PAS 98.2.26  
ABBASI\_MADANI [DZAREBFIS] ; from "FIS"  
ABD-AL-LLAH\_AL-KHATIB [JORGOV] ;\*\*\* 06 Feb 2004  
ABDALLAH\_BAALI [DZAGOV]  
ABDALLAH\_DJABALLAH [DZAOPP] ; PAS 98.2.27the leader of the Reform Movement  
ABDEL\_BOUMEZBEUR [DZAREB] ; from ARMED ISLAMIC GROUP, GIA  
ABDEL\_GHANI\_MESKINI [DZAREB] ; from ARMED ISLAMIC GROUP, GIA  
ABDEL\_HALIM\_KHADDAM [SYRGOV] ;LRP 08 Mar 2004  
ABDEL-HAK\_AYADIA [DZAREB] ; from ARMED ISLAMIC GROUP, GIA  
ABDEL-HALIM ABU GHAZALA [EGYGOV <890501] [---] ;Shan 2/21/93  
ABDELAZIZ\_BELKHADEM [DZAGOV >000826] [DZAELI]  
ABDELAZIZ\_BOUTEFLIKA [DZAGOV >990415] [DZAOPP]  
ABDELAZIZ\_BOUTEFLIKA\_OF\_ALGERIA [DZAGOV >990415] [DZAOPP] ;\*\*\* 06 Feb 2004  
ABDELAZIZ\_ZIARI [DZAGOV] ;\*\*\* 06 Feb 2004  
ABDELHAK\_BENHAMOUDA [DZALAB] ; PAS 98.2.26 Assassinated union leader  
ABDELHAMID\_TEMMAR [DZAGOV >010531] [DZAELI]  
ABDELKADER\_BENSALAH [DZAGOV] ; PAS 98.2.26 speaker of the government-appointed Transitional National Council a quasi-parliament  
ABDELKADER\_HACHANI [DZAOPP] ; Peaceful Society  
ABDELLATIF\_BENACHENHOU [DZAGOV] ;ab 10 Jun 2003  
ABDELMADJID\_SID\_SAID [DZALAB]  
ABDELMAJID\_DAHOU MANE [DZAREB] ; from ARMED ISLAMIC GROUP, GIA  
ABDELMALEK\_SELLAL [EGYGOV] ;\*\*\* 06 Feb 2004  
ABDESSELAM\_BOUCHOUAREB [DZAGOV 010531-020616] [DZAELI]  
ABDUL\_KARIM\_QASIM [###] ;mj 18 Apr 2006  
ABDUL-KARIM\_AL-KABARITI [JORGOV <970307] [JORELI] ;sms 14 Sep 2007

# CountryCodes

- <Country>
- <CountryCode>ATG</CountryCode>
- <CountryName>ANTIGUA\_AND\_BARBUDA</CountryName>
- <COW-Alpha>AAB</COW-Alpha>
- <COW-Numeric>58</COW-Numeric>
- <FIPS-10>AC</FIPS-10>
- <ISO3166-alpha2>AG</ISO3166-alpha2>
- <ISO3166-alpha3>ATG</ISO3166-alpha3>
- <Nationality>ANTIGUANS</Nationality>
- <Nationality>BARBUDANS</Nationality>
- <Nationality>ANTIGUA</Nationality>
- <Nationality>BARBUDA</Nationality>
- <Capital>SAINT\_JOHN'S</Capital>
- <Capital>ST.\_JOHN'S</Capital>
- <MajorCities>
- REDONDA
- </MajorCities>
- <Premiers>
- VERE\_CORNWALL\_BIRD\_ [19670227 - 19710214] [19760201 - 19811101]
- GEORGE\_WALTER\_ [19710214 - 19760201] [B:19280101] [D:20080101]
- </Premiers>
- <Governors-GENERAL>; representing the british monarch as head of state
- SIR\_WILFRED\_E.\_JACOBS\_ [19811101 - 19930610]
- JAMES\_CARLISLE\_ [19930610 - 20070717] [B:19370101]
- LOUISE\_LAKE-TACK\_ [20070717] [B:19440101]
- </Governors-GENERAL>
- <Prime ministers>
- VERE\_CORNWALL\_BIRD\_ [19811101 - 19940309]
- LESTER\_BIRD\_ [19940309 - 20040324] [B:19380101]
- BALDWIN\_SPENCER\_ [20040324] [B:19480101]
- </Prime ministers>
- </Leaders>
- <Government>
- <Synonyms>

# PoliNER/CodeCatcher

# CodeCatcher

SENTENCES: [ N = 743 ]

- [ 1 ] { Israeli }[ISR] shelling of a central { Gaza }[PSE] stronghold of the Ezzedine Al-Qassam Brigades , the armed wing of >>> Hamas <<< , killed one fighter on Tuesday , { Palestinian }[PSE] medics and witnesses said .
- [ 2 ] Top >>> Hamas <<< official Mahmud Al-Zahar on Tuesday rejected the conditions set by { Palestinian }[PSE] president Mahmud Abbas for talks aimed at halting the factional struggle that has torn the territories apart .
- [ 3 ] { Palestinian }[PSE] president Mahmud Abbas said on Monday that he was ready to " open a new page " with >>> Hamas <<< if the Islamist movement gave up its control of the { Gaza }[PSE] Strip .
- [ 4 ] { Israeli }[ISR] forces in the northern { Gaza }[PSE] Strip killed seven { Palestinians }[PSE] overnight , including three militants from the >>> Hamas <<< movement that has ruled the { Gaza }[PSE] Strip since June , medics said on Wednesday .
- [ 5 ] Former { Palestinian }[PSE] { lawmakers }[~LEG] and journalists shaved their heads in public on Wednesday in protest at the humiliating shaving of the moustache of a Fatah official by >>> Hamas <<< men in { Gaza }[PSE] .

SENTENCE:

{ Israeli }[ISR] shelling of a central { Gaza }[PSE] stronghold of the Ezzedine Al-Qassam Brigades , the armed wing of >>> Hamas <<< , killed one fighter on Tuesday , { Palestinian }[PSE] medics and witnesses said .

CODES: 1:ISR 2:PSE

:: Hamas\_ [ PSE ]

->□

# CodeCatcher

## Code

A maximum of nine codes are extracted from the selected sentence; these are displayed after the selected sentence. The initial code defaults to the rst code prior to the target if it exists, otherwise the first code after the target. If there are no codes, the code is set to the discard code ###.

c< n > : add the code < n > from the list

c< n >+< m >: combine codes < n > and < m >. If the second code is an agent code|i.e.

starts with |the is removed.

c+< n >: append code < n >

ca: append text to the code

cr: remove the last three characters from the code

cc: clear code (set to empty string)

cn: enter a complete new code

c-: set null code [- - -]

c-< n; a; s; t;# >: set codes [NOUN],[ADJT],[STOP],[TIME],[NUMR]

c#: set simple discard code ###

c#< s; h; f; a >: set discard codes for sports, history, formatting, arts

# Using Multiple News Sources

- “Splicing” events from multiple sources has proven to be a major problem. Coverage of major events appears to be similar but coverage of minor events such as meetings and comments can vary substantially
- All sources appear subject to “media fatigue”—interest in a crisis will decline over time. Competition between stories—for example US-Iraq versus Israel-Palestine—also affects coverage.
- Overall availability of machine-readable news stories increased dramatically starting around 2000, creating a serious discontinuity. Stories continue to increase exponentially

# Major WEIS Categories

01	Yield	11	Reject
02	Comment	12	Accuse
03	Consult	13	Protest
04	Approve	14	Deny
05	Promise	15	Demand
06	Grant	16	Warn
07	Reward	17	Threaten
08	Agree	18	Demonstrate
09	Request	19	Reduce Relationship
10	Propose	20	Expel
		21	Seize
		22	Force



# Goldstein Scale for WEIS Events

010: [1.0] YIELD  
011: [0.6] SURRENDER  
012: [0.6] RETREAT  
013: [2.0] RETRACT  
014: [3.0] ACCOMODATE, CEASEFIRE  
015: [5.0] CEDE POWER

020: [0.0] COMMENT  
021: [-0.1] DECLINE COMMENT  
022: [-0.4] PESSIMISTIC COMMENT  
023: [-0.2] NEUTRAL COMMENT  
024: [0.4] OPTIMISTIC COMMENT

070: [7.0] REWARD  
071: [7.4] EXTEND ECON AID  
072: [8.3] EXTEND MIL AID  
073: [6.5] GIVE OTHER ASSISTANCE

110: [-4.0] REJECT  
111: [-4.0] TURN DOWN  
112: [-4.0] REFUSE  
113: [-5.0] DEFY LAW

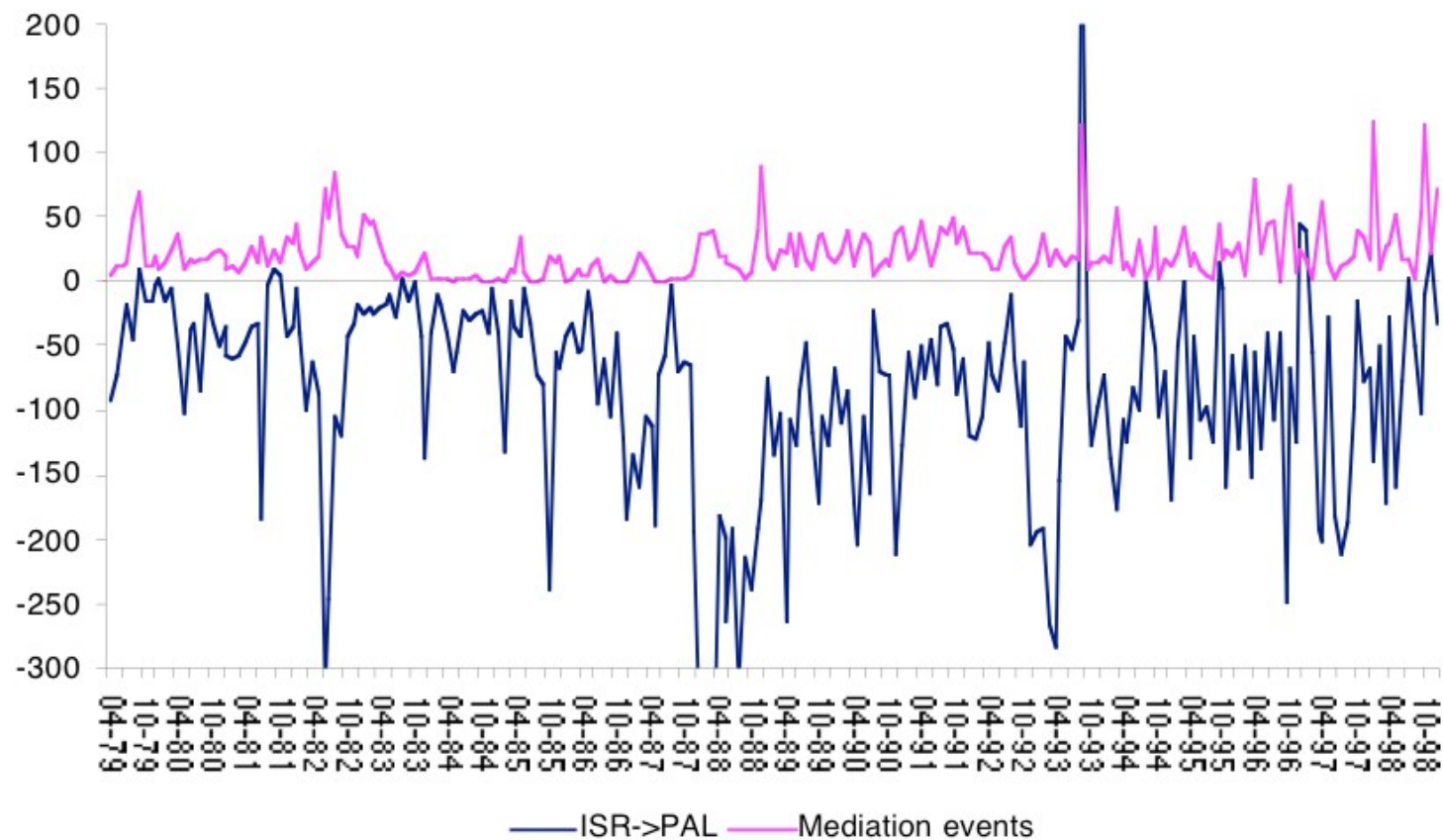
170: [-6.0] THREATEN  
171: [-4.4] UNSPECIFIED THREAT  
172: [-5.8] NONMILITARY TRHEAT  
173: [-7.0] SPECIFIC THREAT  
174: [-6.9] ULTIMATUM

220: [-9.0] FORCE  
221: [-8.3] NONINJURY DESTRUCTION  
222: [-8.7] NONMIL DESTRUCTION  
223: [-10.0] MILITARY ENGAGEMENT

# Event Categories

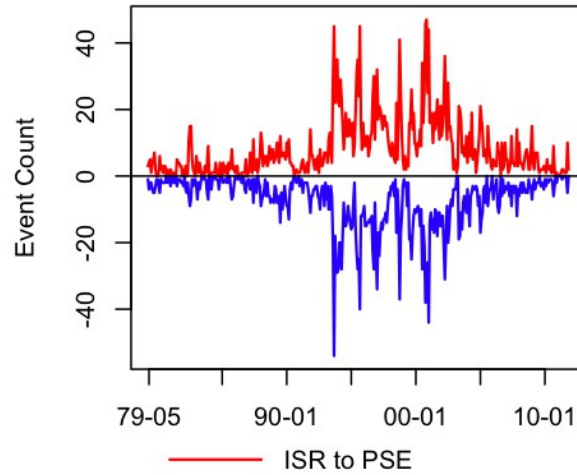
- Verbal Cooperation: The occurrence of dialogue-based meetings (i.e. negotiations, peace talks), statements that express a desire to cooperate or appeal for assistance (other than material aid) from other actors. CAMEO categories 01 to 05.
- Material Cooperation: Physical acts of collaboration or assistance, including receiving or sending aid, reducing bans and sentencing, etc. CAMEO categories 06 to 09.
- Verbal Conflict: A spoken criticism, threat, or accusation, often related to past or future potential acts of material conflict. CAMEO categories 10 to 14.
- Material Conflict: Physical acts of a conflictual nature, including armed attacks, destruction of property, assassination, etc. CAMEO categories 15 to 20.

# Israel-Palestine: Conflict and mediation 1979-98

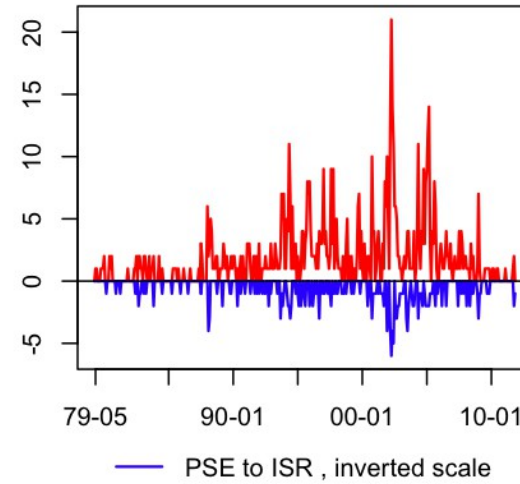


# Levant Event Data Set

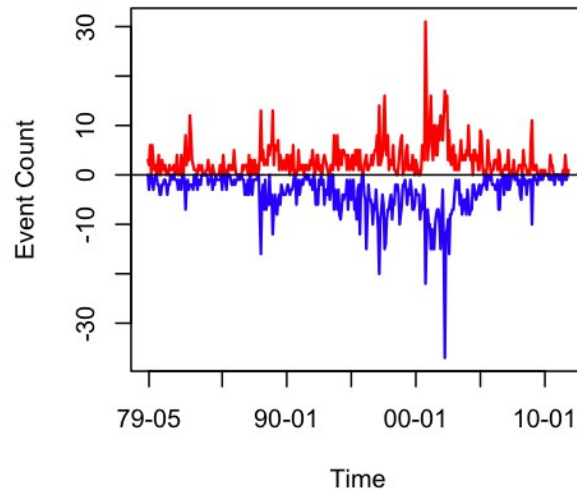
**Verbal Cooperation**



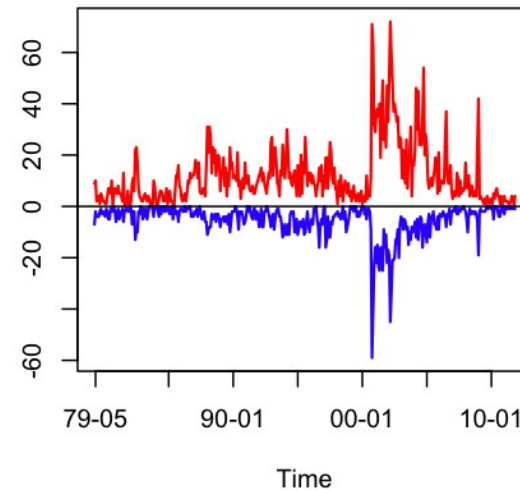
**Material Cooperation**



**Verbal Conflict**

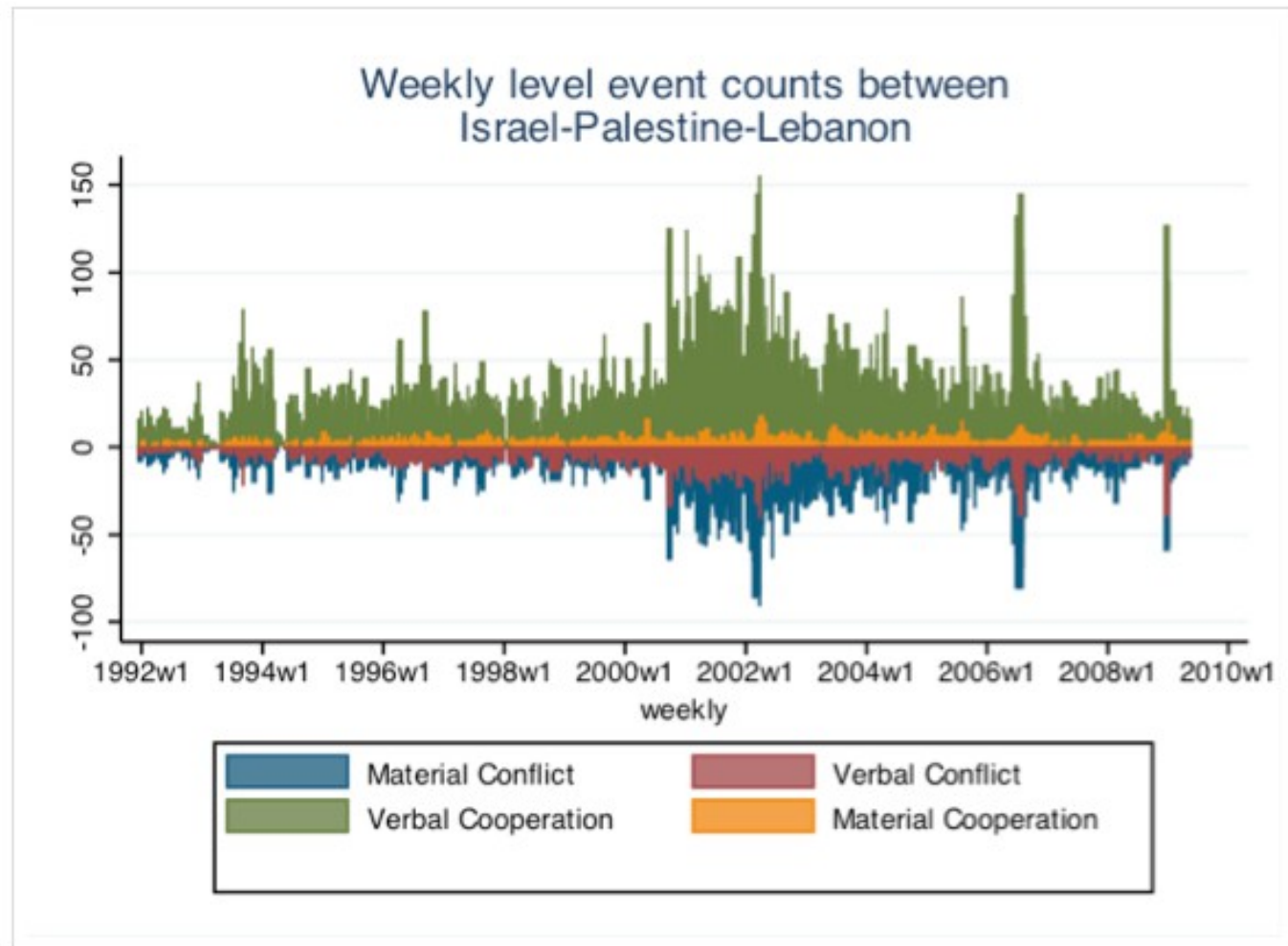


**Material Conflict**

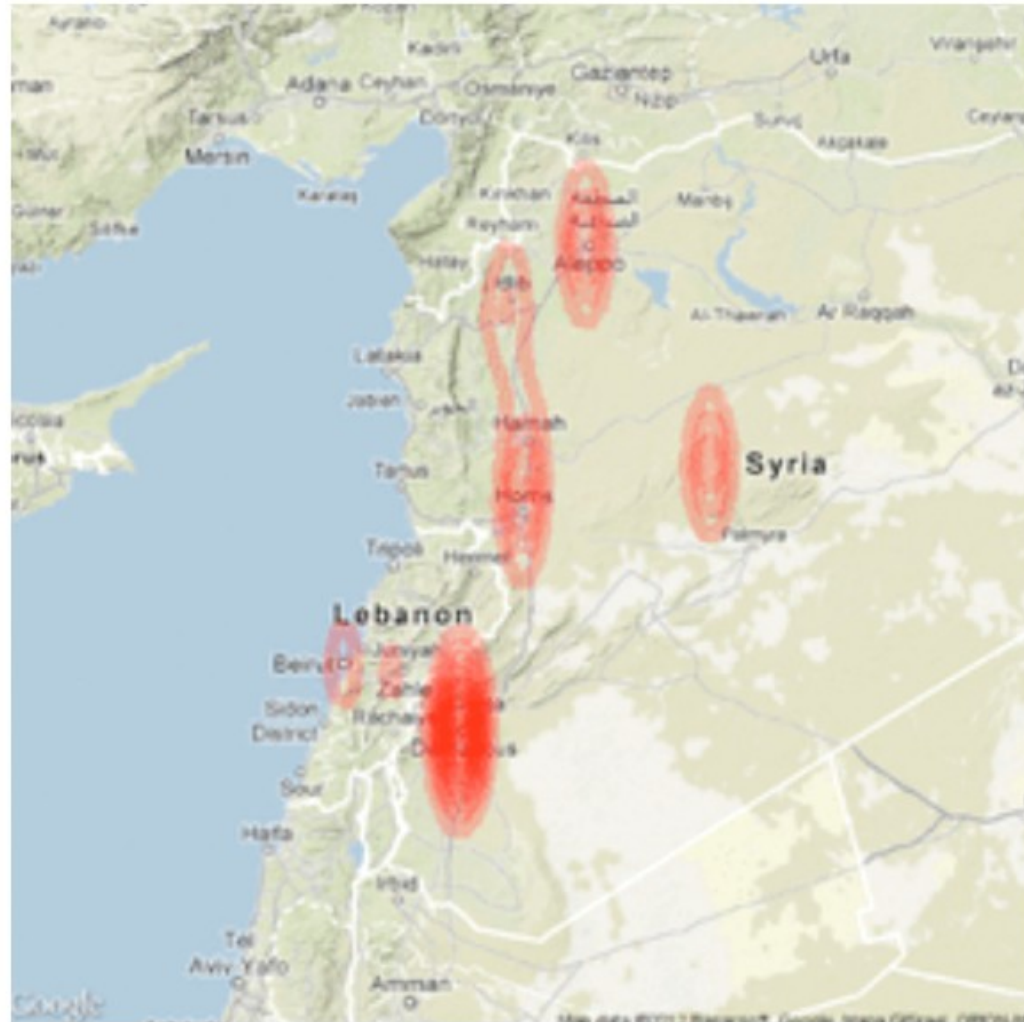


# Israel-Palestine 1992-2010

Visualization by Jay Yonamine

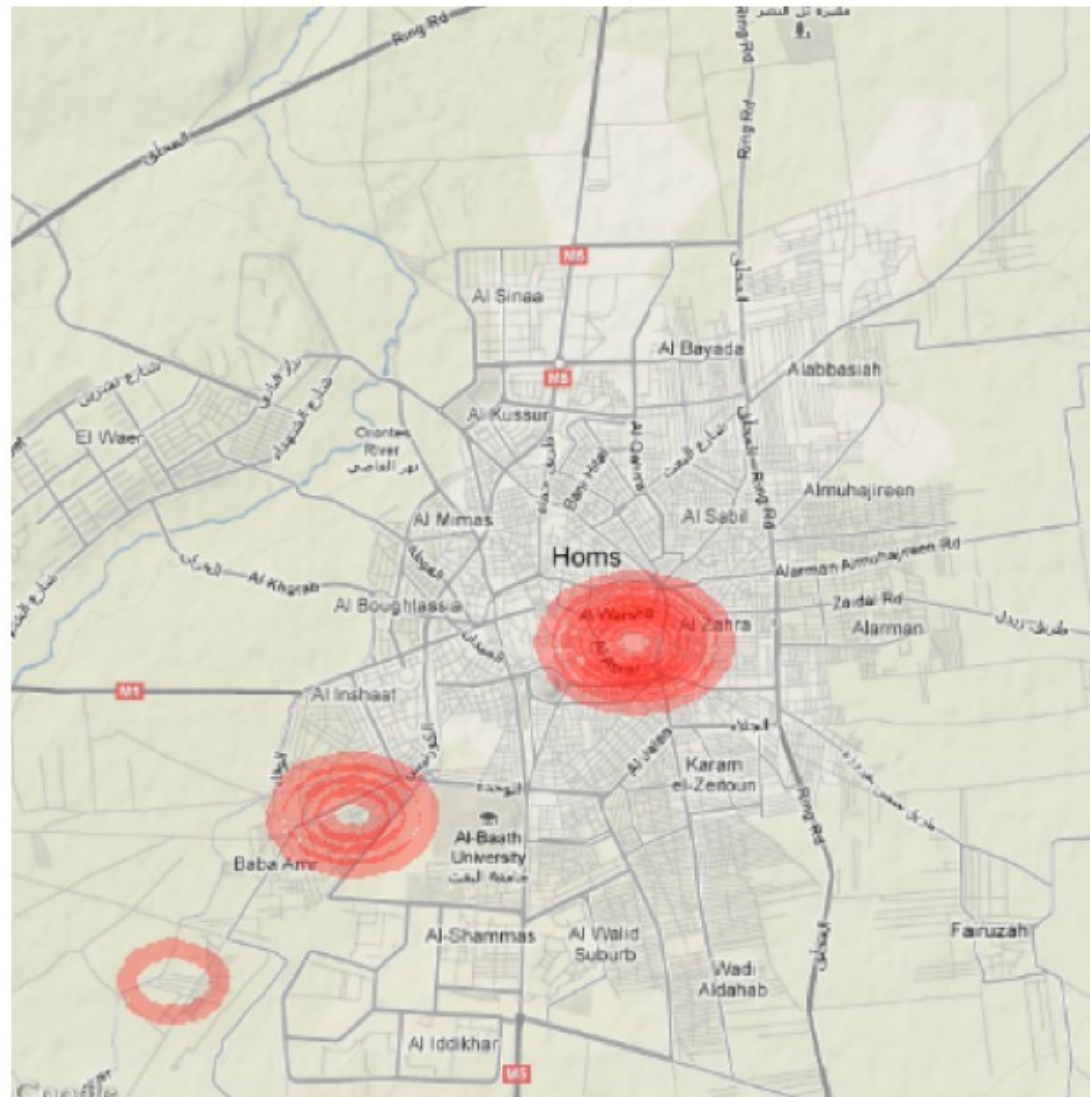


# Syria, April 2012

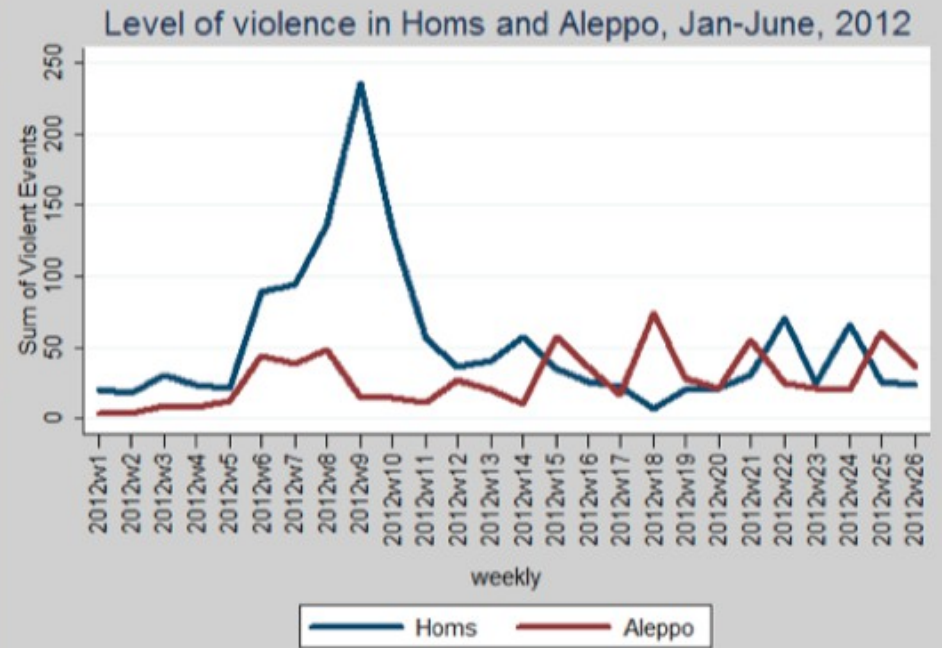
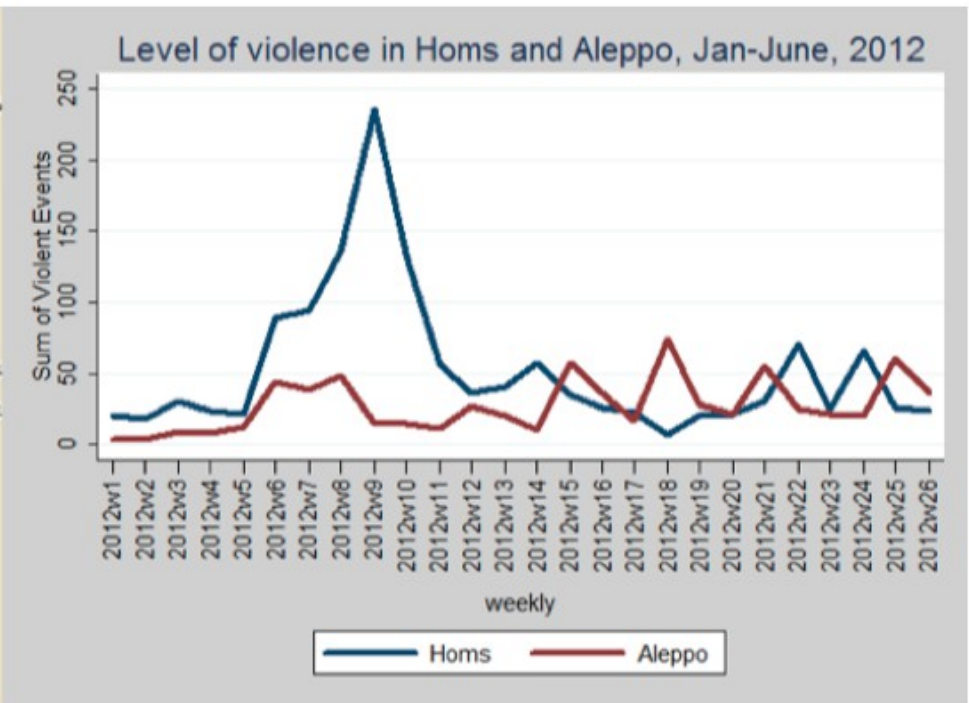




# Violence in Homs, Jan to June 2012



~~~~~

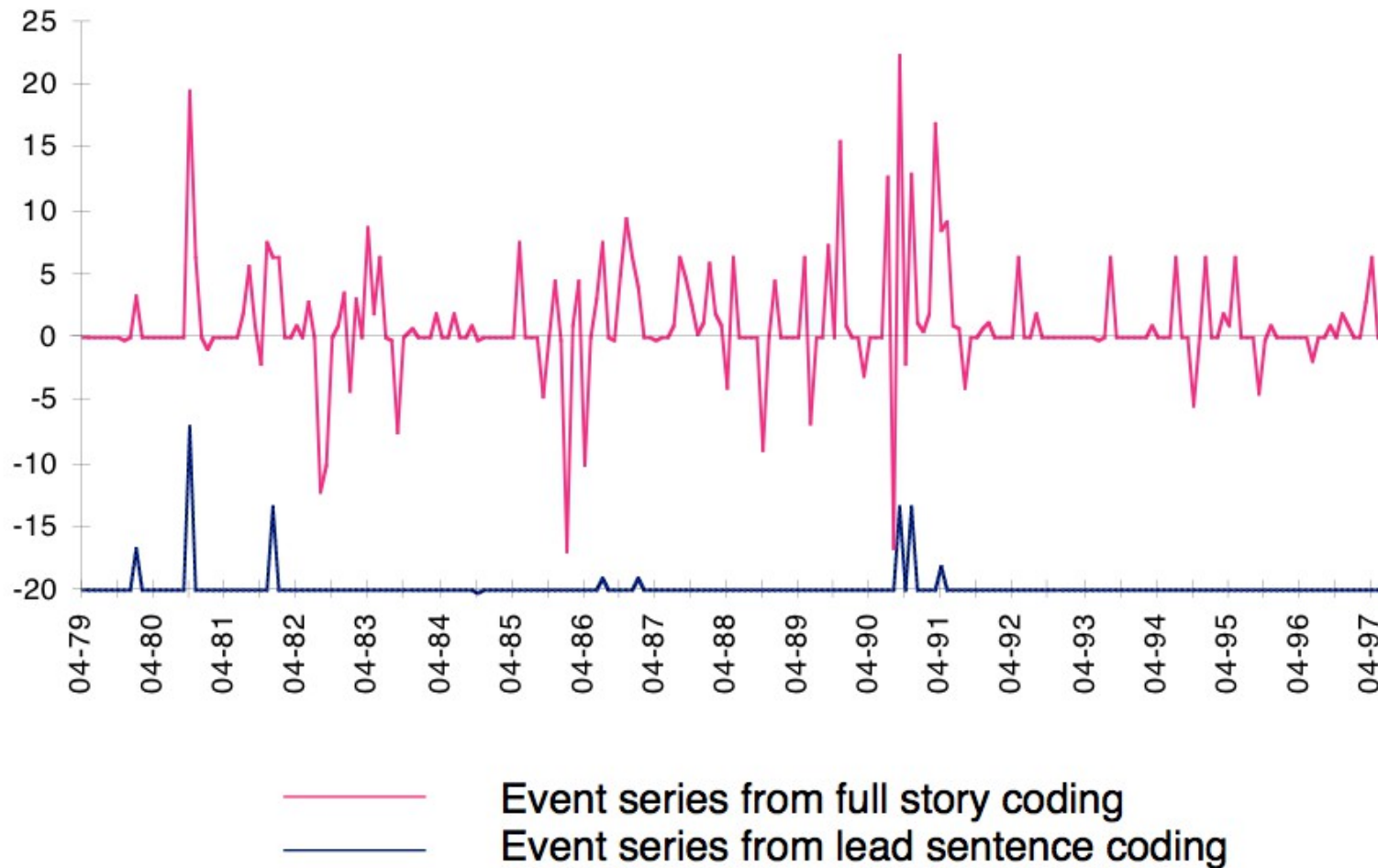




# Goldstein series: UAE-Kuwait, 1979-97

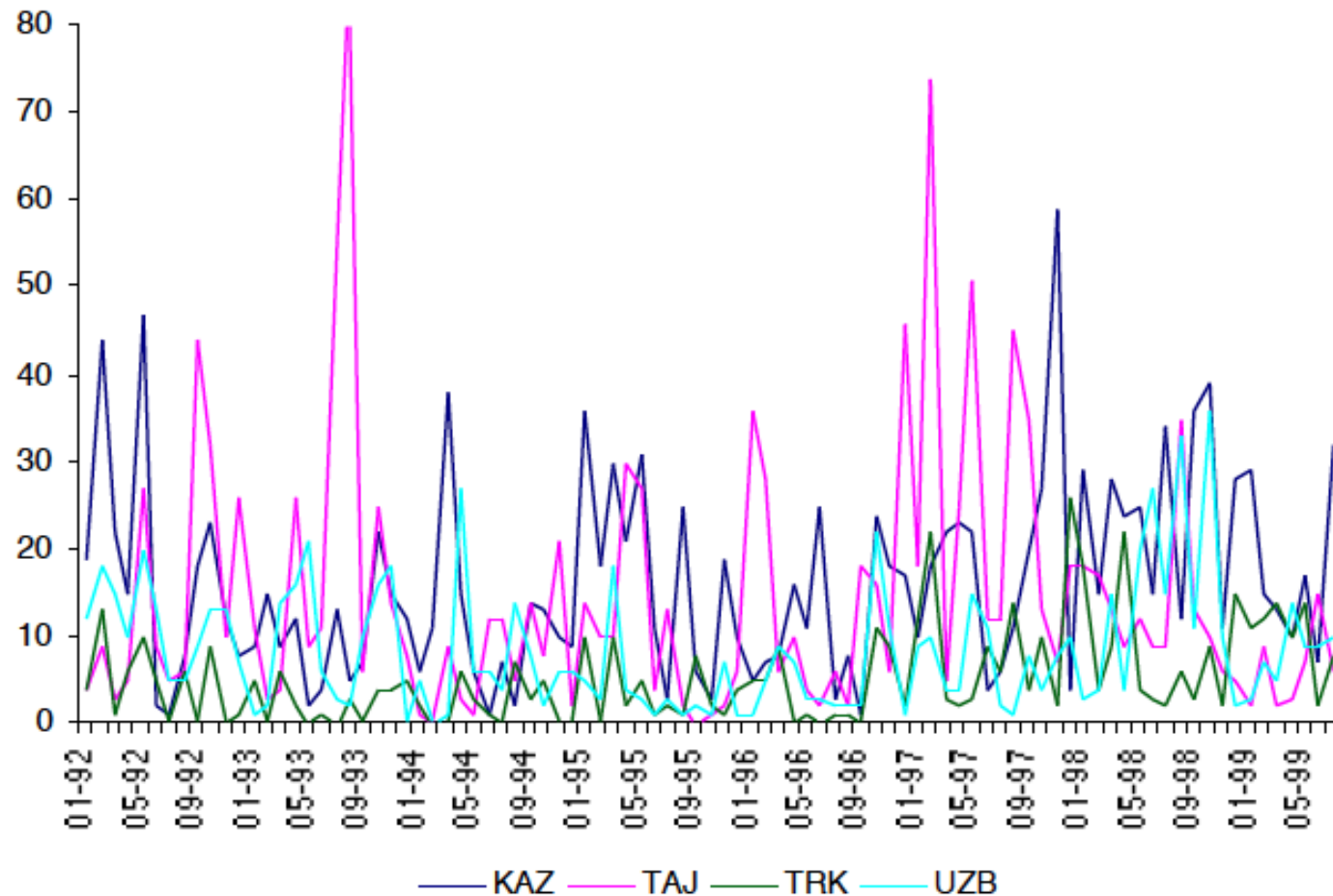
## Full-story vs. lead-sentence events

UAE > KUW



# Worst graph ever produced by the KEDS project

**Figure 3.11. Number of events generated per month for Kazakhstan, Tajikistan, Turkmenistan and Uzbekistan, 1992-1999**



# Hidden Markov models: Accuracy by positive and negative predictions

- “Correct”—percentage of the weeks that were correctly forecast, the percentage of time that a high or low conflict week would have been predicted correctly.
- “Forecast”—percentage of the weeks that were forecast as having high or low conflict actually turned out to have the predicted characteristic; the percentage of time that a type of prediction is accurate.

# Balkans Hidden Markov Model: Accuracy for 23-Category Coding System

| Experiment | %accuracy | % high<br>correct | % low<br>correct | % high<br>forecast | % low<br>forecast |
|------------|-----------|-------------------|------------------|--------------------|-------------------|
| P1         | 77.6      | 29.3              | 89.5             | 40.8               | 83.7              |
| P3         | 76.0      | 29.0              | 87.9             | 37.9               | 82.9              |
| P6         | 76.9      | 25.9              | 90.6             | 42.6               | 82.0              |
| N1         | 54.2      | 92.7              | 45.3             | 28.1               | 96.4              |
| N3         | 49.0      | 88.1              | 39.6             | 25.9               | 93.3              |
| N6         | 47.7      | 88.5              | 37.4             | 26.3               | 92.8              |

# Balkans Hidden Markov Model: Accuracy for 5-Category Coding System

| Experiment | %accuracy | % high<br>correct | % low<br>correct | % high<br>forecast | % low<br>forecast |
|------------|-----------|-------------------|------------------|--------------------|-------------------|
| P1         | 74.4      | 46.2              | 81.5             | 38.9               | 85.6              |
| P3         | 71.7      | 44.1              | 78.9             | 35.4               | 84.4              |
| P6         | 71.4      | 44.2              | 78.8             | 36.4               | 83.8              |
| N1         | 61.9      | 90.7              | 54.6             | 33.7               | 95.8              |
| N3         | 57.8      | 87.0              | 50.2             | 31.4               | 93.6              |
| N6         | 56.8      | 85.9              | 48.8             | 31.5               | 92.7              |

# Difference in Accuracy between 23-Category and 5-Category Coding Systems

| Experiment | % accuracy | % high correct | % low correct | % high forecast | % low forecast |
|------------|------------|----------------|---------------|-----------------|----------------|
| P1         | 3.2        | -16.9          | 8.0           | 1.9             | -1.9           |
| P3         | 4.3        | -15.1          | 9.0           | 2.5             | -1.5           |
| P6         | 5.5        | -18.3          | 11.8          | 6.2             | -1.8           |
| N1         | -7.7       | 2.0            | -9.3          | -5.6            | 0.6            |
| N3         | -8.8       | 1.1            | -10.6         | -5.5            | -0.3           |
| N6         | -9.1       | 2.6            | -11.4         | -5.2            | 0.1            |

Positive value: 23-category has higher accuracy

# Simplifying Event Scales

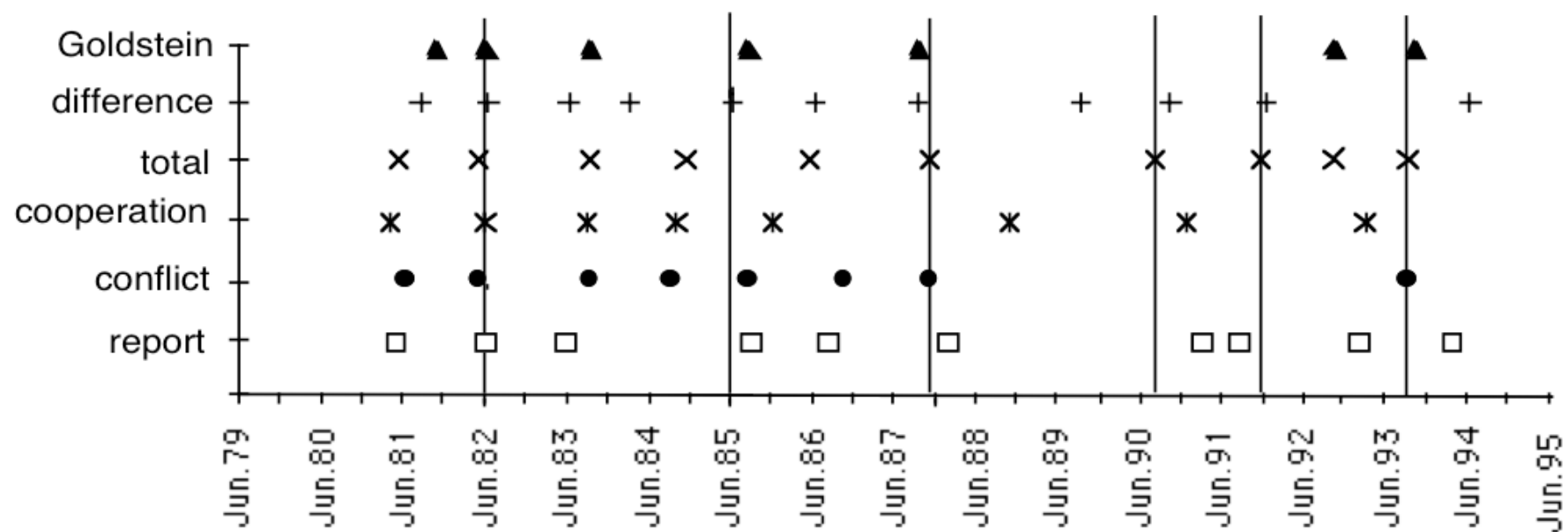
Goldstein: Goldstein weights  
difference: cooperative events = 1; conflictual events = -1.  
total: all events = 1.  
conflict: cooperative event = 0; conflictual events = 1.  
cooperation: cooperative event = 1; conflictual events = 0.  
report: 1 if any event was reported in the month, 0  
otherwise

# Discriminant Analysis Results

| Weighting scheme | %correct | variance explained | canonical correlation | Wilks' $\lambda$ | significance | # factors |
|------------------|----------|--------------------|-----------------------|------------------|--------------|-----------|
| Goldstein        | 85.6%    | 76.3%              | 0.85                  | 0.008            | <.001        | 6         |
| difference       | 89.7%    | 74.7%              | 0.85                  | 0.007            | <.001        | 7         |
| total            | 94.4%    | 83.0%              | 0.93                  | 0.001            | <.001        | 6         |
| conflict         | 88.2%    | 76.9%              | 0.86                  | 0.007            | <.001        | 6         |
| cooperation      | 92.3%    | 82.2%              | 0.91                  | 0.002            | <.001        | 7         |
| report           | 89.2%    | 73.6%              | 0.87                  | 0.008            | <.001        | 7         |
| random date      | 61.0%    | 69.5%              | 0.66                  | 0.131            | .37          | 0         |
| random dyad      | 57.4%    | 68.8%              | 0.67                  | 0.119            | .18          | 0         |



# Cluster boundaries under various weighting systems



# Why does detailed coding make so little difference?—sources of error in event data

- Reporting error
  - Missing events—limited reporting, censorship
  - False events—rumors and propaganda
- Coding error
  - Individual—coders are not correctly implementing the event coding system
  - Systemic—event coding system does not reflect political behavior
- Model specification
  - model may be using the wrong indicators
  - mathematical structure of the model does not produce good predictions
  - Models with diffuse information structures—neural networks, VAR, HMM—are good at adapting to missing information

# Event Categories

- Verbal Cooperation: The occurrence of dialogue-based meetings (i.e. negotiations, peace talks), statements that express a desire to cooperate or appeal for assistance (other than material aid) from other actors. CAMEO categories 01 to 05.
- Material Cooperation: Physical acts of collaboration or assistance, including receiving or sending aid, reducing bans and sentencing, etc. CAMEO categories 06 to 09.
- Verbal Conflict: A spoken criticism, threat, or accusation, often related to past or future potential acts of material conflict. CAMEO categories 10 to 14.
- Material Conflict: Physical acts of a conflictual nature, including armed attacks, destruction of property, assassination, etc. CAMEO categories 15 to 20.

# ICEWS Substate Actor Categories

- gov: government agents such as the executive, police, and military
- par: political parties
- opp: armed opposition---rebels and military groups
- soc: society in general---civilians, businesses, professional groups
- ios: international actors
- usa: United States

# High-volume, near-real-time coding

- News sources from RSS feeds, news aggregators and other web-based sources
- These are relatively stable, but the formatters and downloading still need occasional updating
- Background processing for
  - Parsing
  - Location
  - Context
  - New actor/entity identification
  - Duplicate reports
- Recoding in cluster computing environment
  - ICEWS: 9-million stories can be recoded in about half an hour using a 12-node cluster

# KEDS High-volume processing suite

Objective: “embarrassingly parallel” processing on a cluster computer (i.e. just split and recombine the files; similar to Google's MapReduce)

## Tasks

- Sort multiple files into chronological order
- Duplicate detection
- Post-download filtering (particularly if using LN) and date-consistency checking (eliminates records with bad or out-of-range dates)
- Create *.project* files which run with current version of TABARI in parallel mode
- Split file for coding, then recombine results into a single chronological sequence

# Pre-parsing

- Use open-source linguistics tools—not the coding program—to handle most of the parsing tasks.

- OpenNLP: <http://opennlp.apache.org>

GATE: <http://gate.ac.uk/>

University of Illinois Cognitive Computation Group:

<http://cogcomp.cs.illinois.edu/page/software>

Stanford NLP Group: <http://nlp.stanford.edu/software/index.shtml>

LingPipe's “Competition” page: <http://alias-i.com/lingpipe/web/competition.html>

- Dictionaries would then be modified to use this information

- Parsing tasks

- Entity identification/disambiguation
- Parts of speech, particularly noun/verb disambiguation
- Subject, verb and object phrase delineation
- Pronoun coreferencing

- With sufficient information, coding becomes largely a bookkeeping problem: almost all of the knowledge is in the dictionaries

# Contextual Coding

- Determine the context of the report from the complete story, rather than each individual sentence
- Location
  - Ideally to as much detail as possible, using gazetteers, most in the public domain
  - However, some stories do not have a location
  - Location can also be used to resolve agents
  - Resolves ambiguous common names and acronyms
- Better filtering of sports, business, entertainment and historical stories
- General categories and then the use of specialized dictionaries
  - For example “attack” has a different meaning depending on whether a story involved military action, debate or cyber-attack



# Questions?

Philip A. Schrod  
Political Science  
227 Pond Laboratory  
Pennsylvania State University  
University Park, PA 16801

Phone: 814-863-8978

Email: [schrod@psu.edu](mailto:schrod@psu.edu)

Project Web Site: <http://eventdata.psu.edu>





## Some additional considerations

---

# But Phil, the best models are classified!

- Hollywood tells me so
- Yeah, right...
- No systematic evidence of this: if it is true, government is spending vast resources to obscure this fact
- Clearly isn't operating at the policy level
- Probably some models have worked at some points in the past but they have not proven robust
  - Serious snake-oil sales going on here as well...
- Even if this is true, we need to reverse-engineer these to get them into the unclassified literature and acquaint policy-makers with the techniques
- [but it probably isn't true...]

# What methods do *real* intelligence agencies use to predict political events?

“Two weeks ago, a group of senior intelligence officials in the [U.S.] Defense Dept. sat for an hour listening to a briefing by Michael Drosnin, who claims—I am not making this up—that messages encoded in the Hebrew text of the Old Testament provide clues to the whereabouts of Osama bin Laden. He has given similar briefings to top officials of Mossad, the Israeli intelligence agency.”

Bill Keller

*International Herald-Tribune*, 8-9 March 2003, pg. 6

[Insight as to why the Bush administration never found bin-Laden, perhaps?]



# Easy Problems



# Hard Problems

