Political Event Data Analysis: Automated Coding and Conflict Forecasting

Philip A. Schrodt
Department of Political Science
Pennsylvania State University

schrodt@psu.edu
http://eventdata.psu.edu

Presentation at the Tools for Text Workshop
University of Washington/Seattle
14 June 2010

Outline of presentation

- What is event data?
- Early DARPA event data vs. contemporary methods
- RSS-based real-time monitoring
- Current environment for technical political forecasting models
- Political forecasting with event data
- Where to go next?

Reuters Chronology of 1990 Iraq-Kuwait Crisis - 1

July 17, 1990: RESURGENT IRAQ SENDS SHOCK WAVES THROUGH GULF ARAB STATES

Iraq President Saddam Hussein launched an attack on Kuwait and the United Arab Emirates (UAE) Tuesday, charging they had conspired with the United States to depress world oil prices through overproduction.

July 23, 1990: IRAQ STEPS UP GULF CRISIS WITH ATTACK ON KUWAITI MINISTER

Iraqi newspapers denounced Kuwait's foreign minister as a U.S. agent Monday, pouring oil on the flames of a Persian Gulf crisis Arab leaders are struggling to stifle with a flurry of diplomacy.

July 24, 1990: IRAQ WANTS GULF ARAB AID DONORS TO WRITE OFF WAR CREDITS

Debt-burdened Iraq's conflict with Kuwait is partly aimed at persuading Gulf Arab creditors to write off billions of dollars lent during the war with Iran, Gulf-based bankers and diplomats said.

Reuters Chronology of 1990 Iraq-Kuwait Crisis - 2

July 24, 1990: IRAQ, TROOPS MASSED IN GULF, DEMANDS $25 OPEC OIL PRICE

Iraq's oil minister hit the OPEC cartel Tuesday with a demand that it must choke supplies until petroleum prices soar to $25 a barrel.

July 25, 1990: IRAQ TELLS EGYPT IT WILL NOT ATTACK KUWAIT

Iraq has given Egypt assurances that it would not attack Kuwait in their current dispute over oil and territory, Arab diplomats said Wednesday.

July 27, 1990: IRAQ WARNS IT WON'T BACK DOWN IN TALKS WITH KUWAIT

Iraq made clear Friday it would take an uncompromising stand at conciliation talks with Kuwait, saying its Persian Gulf neighbor must respond to Baghdad's "legitimate rights" and repair the economic damage it caused.

July 31, 1990: IRAQ INCREASES TROOP LEVELS ON KUWAIT BORDER

Iraq has concentrated nearly 100,000 troops close to the Kuwaiti border, more than triple the number reported a week ago, the Washington Post said in its Tuesday editions.

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

WEIS Coding of Tiny Subset of 1990 Iraq-Kuwait Crisis

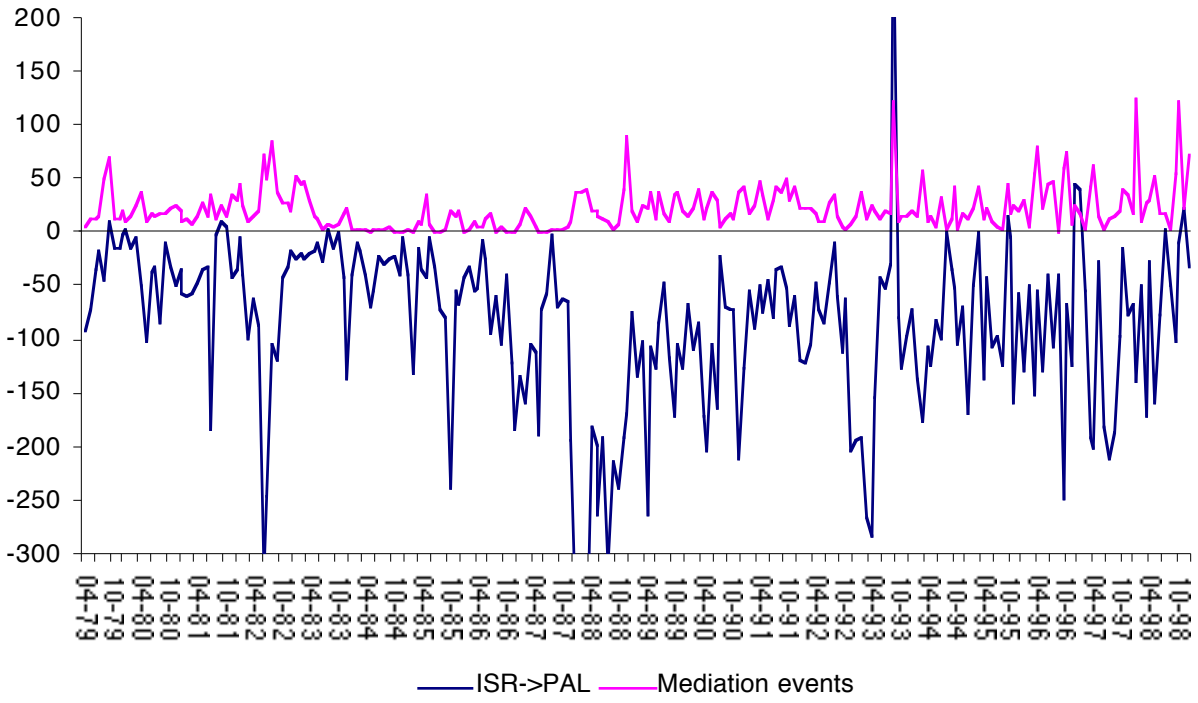
Date	Source	Target	Code	Type of Action
900717	IRQ	KUW	121	CHARGE
900717	IRQ	UAE	121	CHARGE
900723	IRQ	KUW	122	DENOUNCE
900724	IRQ	ARB	150	DEMAND
900724	IRQ	OPC	150	DEMAND
900725	IRQ	EGY	054	ASSURE
900727	IRQ	KUW	160	WARN
900731	IRQ	KUW	182	MOBILIZATION
900801	KUW	IRQ	112	REFUSE
900802	IRQ	KUW	223	MILITARY FORCE

Goldstein Scale for WEIS Events

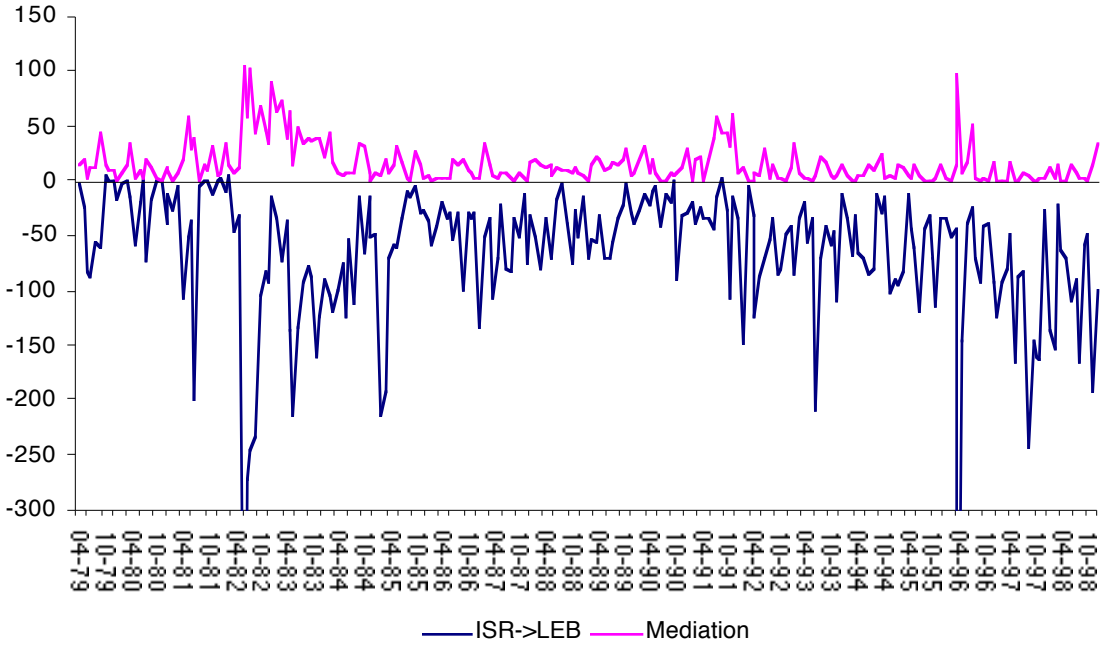
```
010: [1.0]  YIELD
011: [0.6]  SURRENDER
012: [0.6]  RETREAT
013: [2.0]  RETRACT
014: [3.0]  ACCOMODATE, CEASEFIRE
015: [5.0]  CEDE POWER

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
```

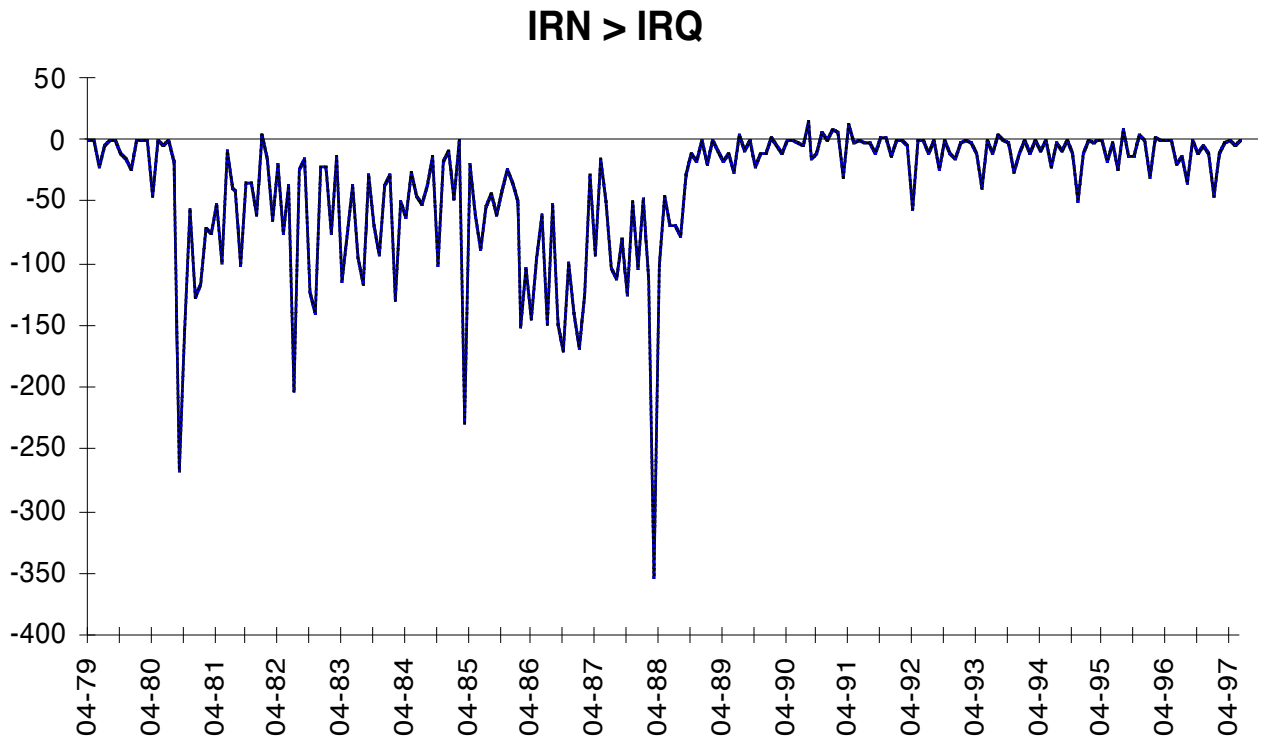
Israel-Palestine: Conflict and mediation 1979-98



Israel-Lebanon: Conflict and mediation 1979-98



Goldstein-scaled series: Iran→ Iraq 1979-97



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.

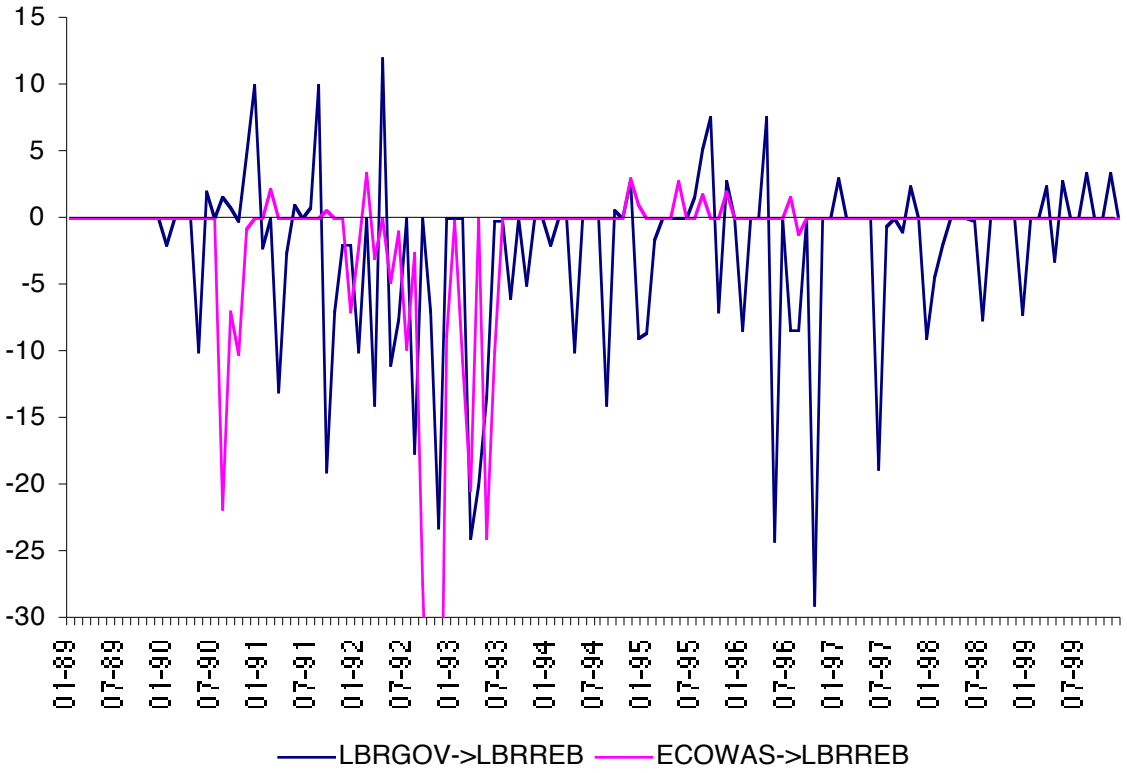
The Turkish attacks in Dohuk Province on Sunday — involving dozens of warplanes and artillery — were the largest known cross-border attack since 2003. They occurred with at least tacit approval from American officials. The Iraqi government, however, said it had not been consulted or informed about the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

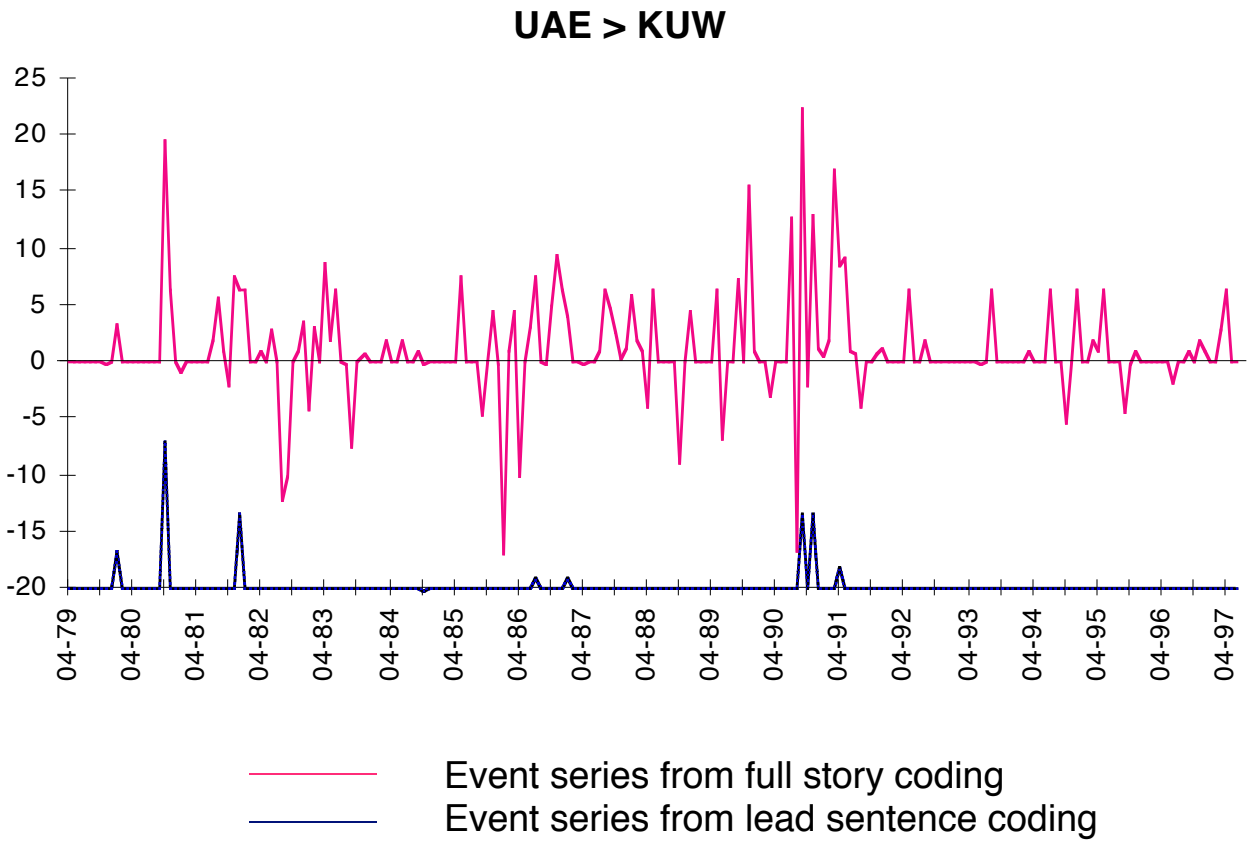
New York Times, 18 December 2007

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

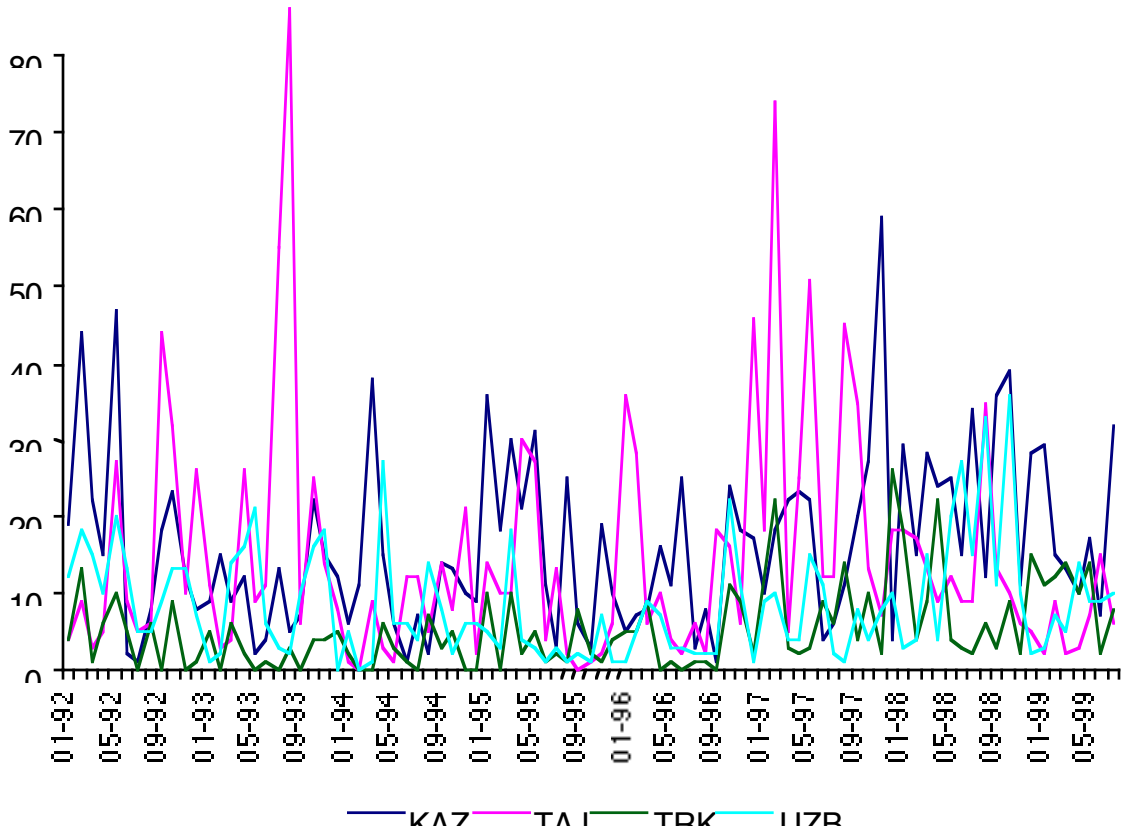
Goldstein series for Liberia and ECOWAS actions toward rebels, 1989-99



Goldstein series: UAE→Kuwait, 1979-97
Full-story vs. lead-sentence events



The worst graphic ever produced by the KEDS project…



What we know about event data in 2010 that *confirms* what we suspected in 1975

- WEIS 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
- It is impossible for human coders to keep up with coding in real time
- Media fatigue is a major factor in event reporting
- Comments and meetings are about 30% to 50% of most event data
- Violent events are reported disproportionately

What we know about event data in 2010 that we *didn't know* in 1975

- Machine coding to a level of accuracy comparable to 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 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 2010 that we *didn't know* in 1975, continued

- News sources vary dramatically in their coverage; these effects differ by region. However, news service reports provide substantially greater coverage than individual newspapers
  - 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

A practical approach to automated coding

Old objective:

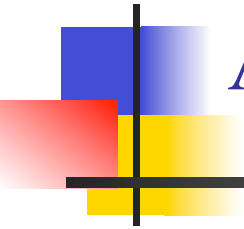
Machine coding should attempt to duplicate human coders

- (which, in fact, can be done: Schrodt and Gerner 1994, Bond et al 1997, Thomas 2001, King and Lowe 2003)

Alternative objective:

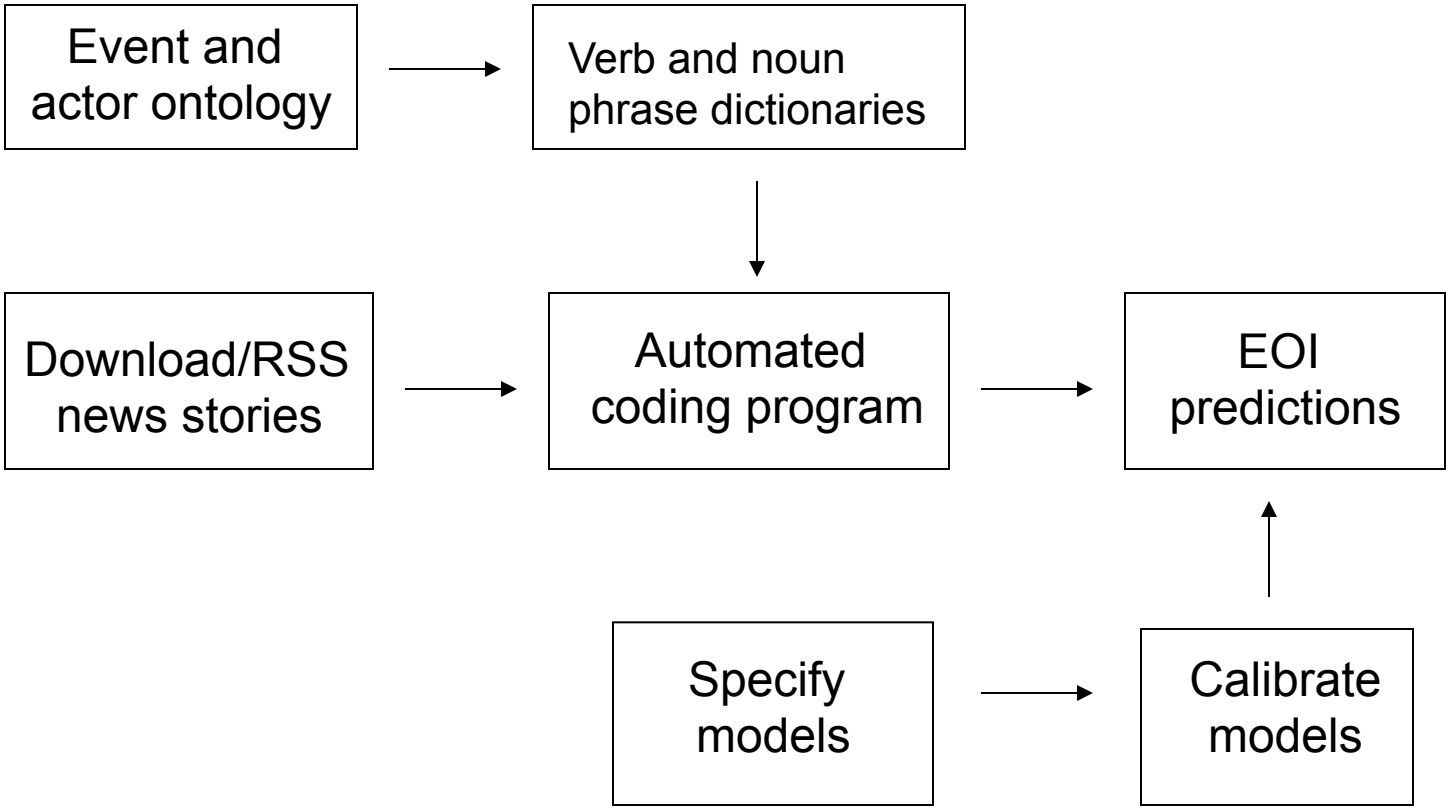
Optimize coding systems and models to use information that can be coded most reliably by machine

More generally, human coding of *event data* is completely irrelevant because human coders can't handle the volume: automated coding is the only practical solution. So, is it good enough?



Automated Event Coding

Event Data Generation Process



Textual Analysis By Augmented Replacement Instructions (TABARI)

- ANSI C++, approximately 14,000 lines of code
- Open-source (GPL)
- Linux, Macintosh and Windows (sort of…) operating systems
- "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
  - Process can be trivially parallelized by splitting the input files, so speed scales almost linearly on a cluster computer
- Standard dictionaries are open source, with around 16,000 verb phrases for events and 8,000 noun phrases for actors (ICEWS: 16,000 noun phrases)

Proprietary Automated Coding Systems

- VRA Coder— full frame-based parser
  - Data set available on Gary King's web site; evaluated by King and Lowe (2003)
- Xenophon—Shellman and Covington, shallow parser
- JABARI-NLP—Lockheed
  - Uses Open-NLP as a pre-parser
- Social Science Automation—probably a shallow parser; based on Profiler+
- BBN, IBM: modifying existing NLP systems for event coding

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 16,000+ verb phrase dictionary

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: 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

Event Model: Core Innovation

- Once calibrated, real-time event forecasting models can be run ***entirely*** without human intervention
  - RSS feeds from news aggregators—Google News, European Media Monitor—provide a rich multi-source flow of news reports in real time
  - These reports can be formatted and coded automatically
  - 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

Components of fully automated system

- Machine coding system
- Machine-compatible event ontologies and dictionaries to implement these
- RSS feeds for news stories
- Automated actor/entity and location detection
- Integration of these parts, typically in an open-source LAMP (Linux, Apache, mySQL, PHP) environment

  http://129.237.60.130/~eventdata/politicalworld.php

RSS Feeds: 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

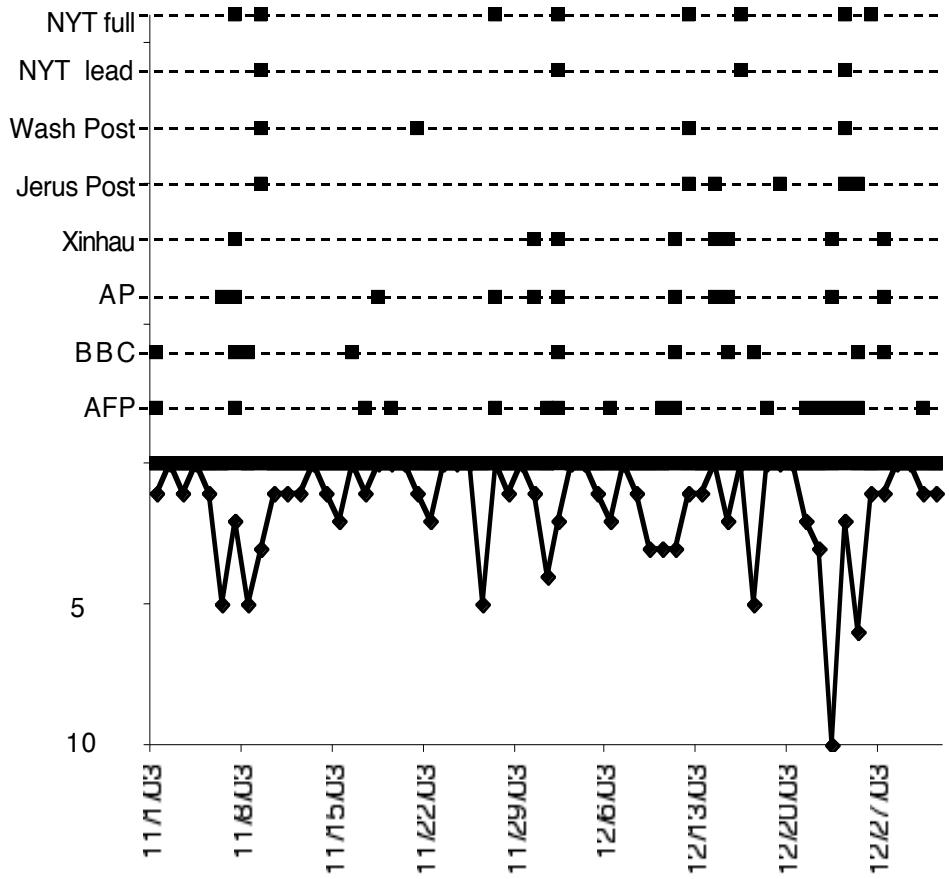
RSS Feeds: Google News

- 4500 English-language sources
- Appears to have facility for duplicate detection
- Multiple channels, including at least two specifically focused on international events
- It's Google…

Number of stories found with "Palestinian killed" NEXIS search string

Newspaper		Wire Service	
Los Angeles Times	3	Xinhau	8
New York Times	4	BBC (Factiva)	10
Washington Post	4	Associated Press	11
Jerusalem Post	6	Agence France Presse	18
New York Times, full text	8		

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



Statistical Forecasting Results

Statistical Forecasting Results

It works in practice if not in theory

Statistical Forecasting Results

It works in practice if not in theory

And by the way, prediction *is* “scientific”

Statistical Forecasting Results

It works in practice if not in theory

And by the way, prediction *is* "scientific"

Losers…

Factors encouraging technical political forecasting

- Conspicuous failures of existing methods
- Success of forecasting models in other behavioral domains
  - Macroeconomic forecasting
  - Elections
  - Demographic and epidemiological forecasting
  - Famine forecasting: USAID FEWS model
  - Example: statistical models for mortgage repayment were quite accurate
- Technological imperative
  - Increased processing capacity
  - Information available on the web
  - "Moore's Law states that computing power doubles every 18 months. Human cognitive ability is pretty much a constant. This leads to some interesting and not always desirable substitution effects"
    Larry Bartels, Princeton University

Factors encouraging technical political forecasting

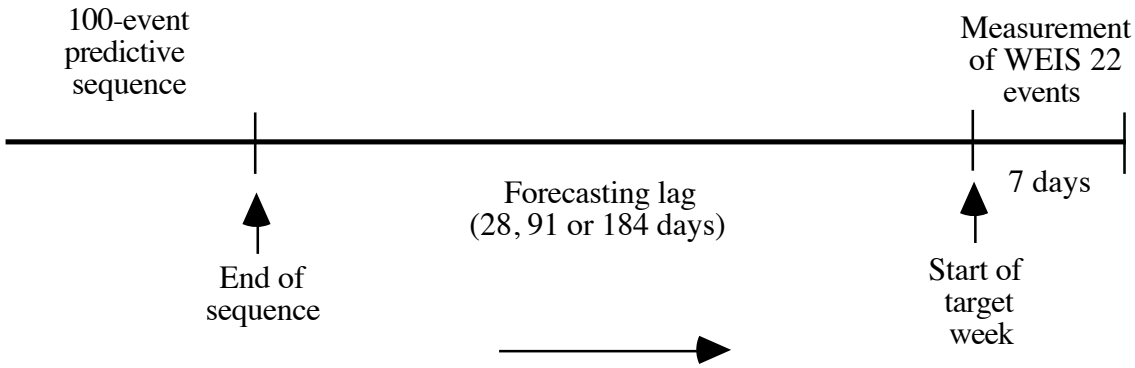
- Demonstrated utility of existing methods
  - Political Instability Task Force
- Decision-makers now expect visual displays of analytical information
  - "They won't read things any more"
- Ahmed Chalabi
  - At least *some* SME sources can be problematic, even if they do understand the language and culture
  - Also see N. Machiavelli (1513) on the topic of trusting exiles

Contemporary Technical Political Forecasting

- State Failures Project 1994-2001
- Joint Warfare Analysis Center 1997
- FEWER [Davies and Gurr 1998]
- Various UN and EU forecasting projects
- Center for Army Analysis 2002-2005
- Swiss Peace Foundation FAST 2000-2006
- Political Instability Task Force 2002-present
- DARPA ICEWS 2007-present

Lead times for policy-relevant forecasting

- Typically 1 to 18 months—6 months is a good guideline
  - Any "prediction" less than 1 month is an autopsy, not a diagnosis
  - Relevant lead time is also a function of the expected intervention —you can deploy a combat team more quickly than you can deploy a division
- Example: prediction scheme used in testing hidden Markov models



Predictive Accuracy with Event Data

- Predictive accuracy using event data models in protracted conflicts is 60% to 80% at policy-relevant lead times using simple statistical methods
- Scaling and high levels of detail have little effect: 50% of the variance is explained by the event report

Linear Regression (r^2) on Material Conflict Event Counts

Lead (Months)	Balkans	Palestine	Lebanon	West Africa
1	0.34	0.45	0.31	0.12
3	0.15	0.29	0.23	0.03 (n.s.)
6	0.06 (.04)	0.27	0.16	0.03 (n.s.)
12	0.04 (n.s.)	0.23	0.16	0.01 (n.s.)

* results are significant at $p<0.0001$ unless otherwise noted. P-value is in (); n.s. = not significant at 0.10 level

Logistic Regression on Event Counts (in sample)

Lead	Balkans	Palestine	Lebanon
50% level			
1 month	73.7%	82.6%	75.3%
6 month	64.3%	74.9%	68.5%
75% level			
1 month	79.6%	79.6%	81.7%
6 month	72.8%	79.2%	75.6%

Logistic Regression on Event Counts (1:3 out-of- sample)

Lead	Balkans	Palestine	Lebanon
50% level			
1 month	64.3%	57.3%	67.7%
6 month	60.1%	--*	56.4%
75% level			
1 month	66.1%	71.0%	82.3%
6 month	61.6%	--	74.6%

*Palestine 6-month forecasts could not be estimated due to insufficient variance in high-conflict data points

Logistic Regression on Event Counts (1:1 out-of- sample)

Lead	Balkans	Palestine	Lebanon
50% level			
1 month	66.7%	64.4%	63.4%
6 month	47.1%	38.1%	46.7%
75% level			
1 month	85.3%	67.8%	75.4%
6 month	87.1%	55.7%	61.3%

Integrated Conflict Early Warning System (ICEWS)

- DARPA funded, around $35-million for Phases I and II
- Focus: 29 countries in Asia; five indicators of political instability, six month forecast window; quarterly data
- First phase involved three competing teams—Lockheed, BBN and SAIC—judged on results in a split-sample test
  - Data was provided for 1997-2004, tested on 2005-2006
- Second phase is Lockheed, only team that passed the benchmarks in Phase I

ICEWS "Events of Interest"

Domestic Political Crisis—Significant opposition to the government, but not to the level of rebellion or insurgency (for example, power struggle between two political factions involving disruptive strikes or violent clashes between supporters)

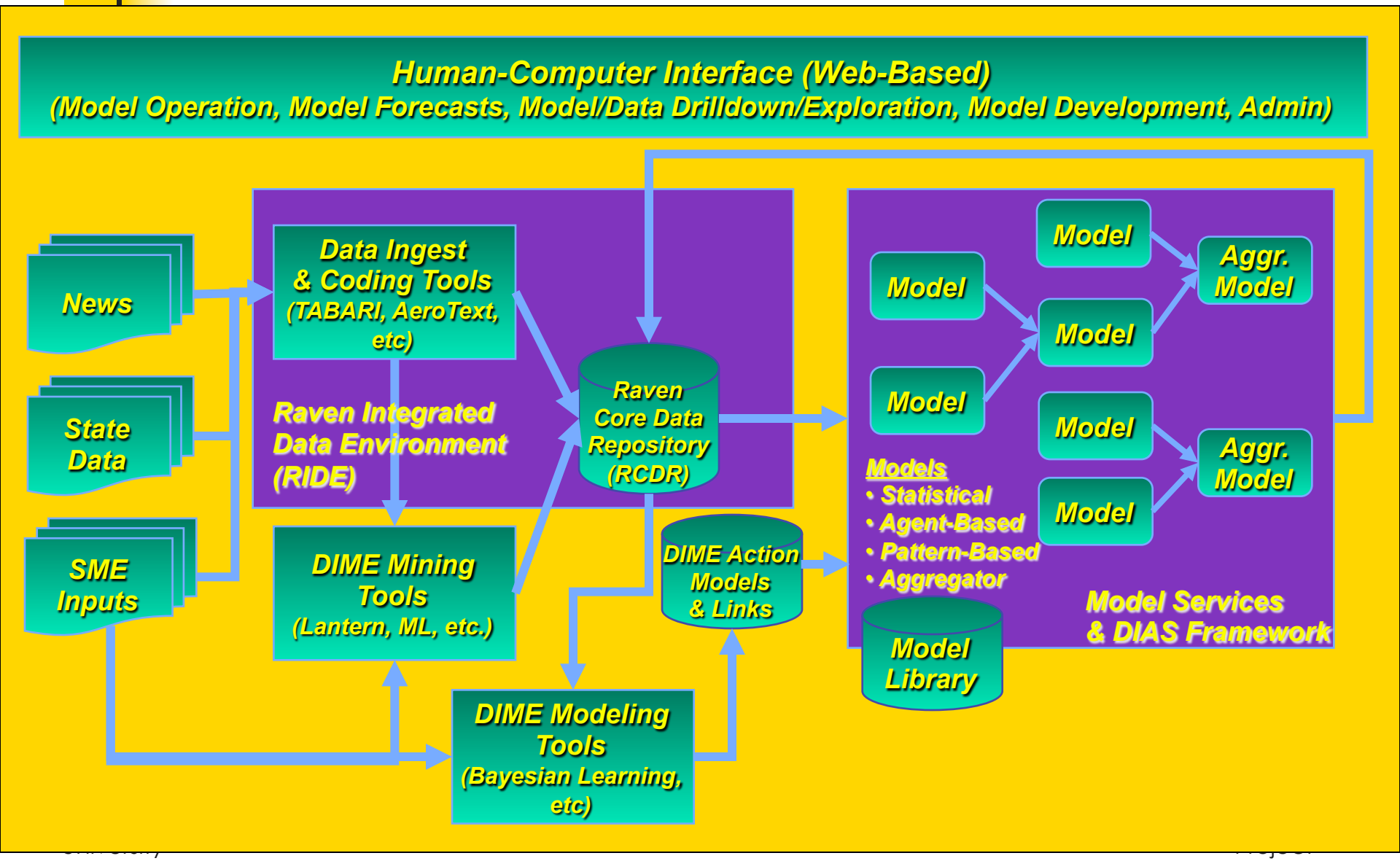
Rebellion—Organized opposition where the objective is to seek autonomy or independence

Insurgency—Organized opposition where the objective is to overthrow the central government

Ethnic/Religious Violence—Violence between ethnic or religious groups that is not specifically directed against the government

International Crisis—Conflict between two or more states or elevated tensions between two or more states that could lead to conflict

Raven Phase 2 Functional View



News & Other Data Sources (Phase 1)

- TABARI open source event data coding tool
  - 6.7M news stories from 75+ sources
  - 253M lines of text
  - 30 dictionaries, 20K entries
  - CAMEO action taxonomy
  - complementing with AeroText in Phase 2
- Country/State data
  - 16+ sources
- SME interviews for agent-based country models

It is estimated that this is the largest automated event coding project to date. Enabled by end-to-end automated process.

As of Jun 2008

		Australia	Bangladesh	Bhutan	Burma (Myanmar)	Cambodia	China	Comoros	Fiji	India	Indonesia	Japan	Korea, North	Korea, South	Laos	Madagascar	Malaysia	Mauritius	Mongolia	Nepal	New Zealand	Papua New Guinea	Philippines	Russia	Singapore	Solomon Islands	Sri Lanka	Taiwan	Thailand	Vietnam
Intnl	BBC World Monitoring Service*	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	The New York Times	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	The Associated Press	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Japan Economic News Wire	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	United Press International	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Inter Press Service	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Agence France Presse	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Xinhua General News Service	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Deutsche Presse Agentur	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Local Sources	Bangkok Post				X	X				X				X		X		X							X				X	X
	Central News Agency- Taiwan					X	X						X	X						X								X		
	India Today		X	X		X			X											X						X				
	The Edge Malaysia				X	X				X				X		X								X				X	X
	The Edge Singapore	X			X	X				X				X		X					X	X		X				X	X
	The Jakarta Post	X			X	X				X				X		X					X	X		X				X	X
	The Nation (Thailand)	X			X	X				X				X		X						X		X				X	X
	The Nation (Pakistan)		X	X			X			X										X						X			
	The Pakistan Newswire		X	X			X			X										X						X			
	The Statesman- India		X	X			X			X										X						X			
SMEs	UPENN Questionaire																													
	IDI Leaders Questionaire																													

Being Purchased or Developed

Acquired

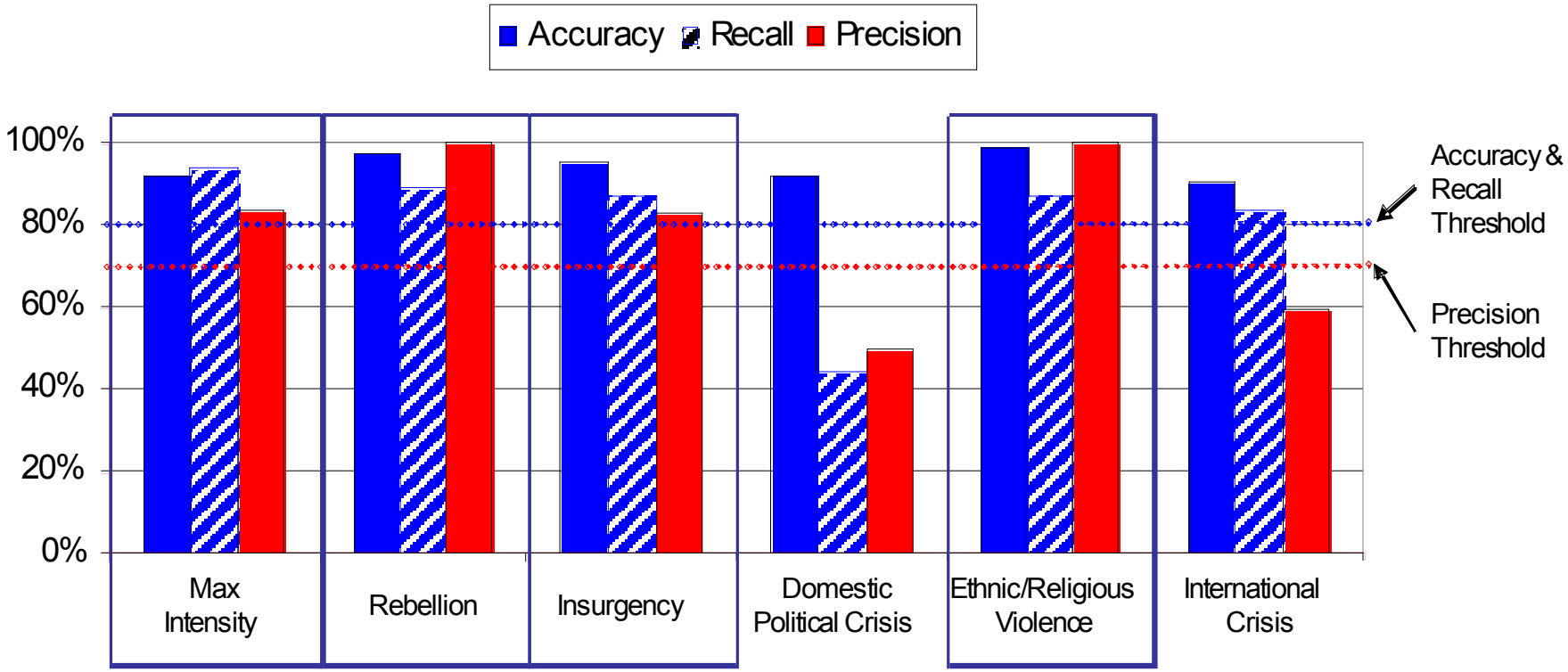
X Encoded (multiple times)

*Operate around the clock to monitor more than 3,000 radio, TV, press, internet and news agency sources, translating from up to 100 languages.

ICEWS Evaluation Criteria

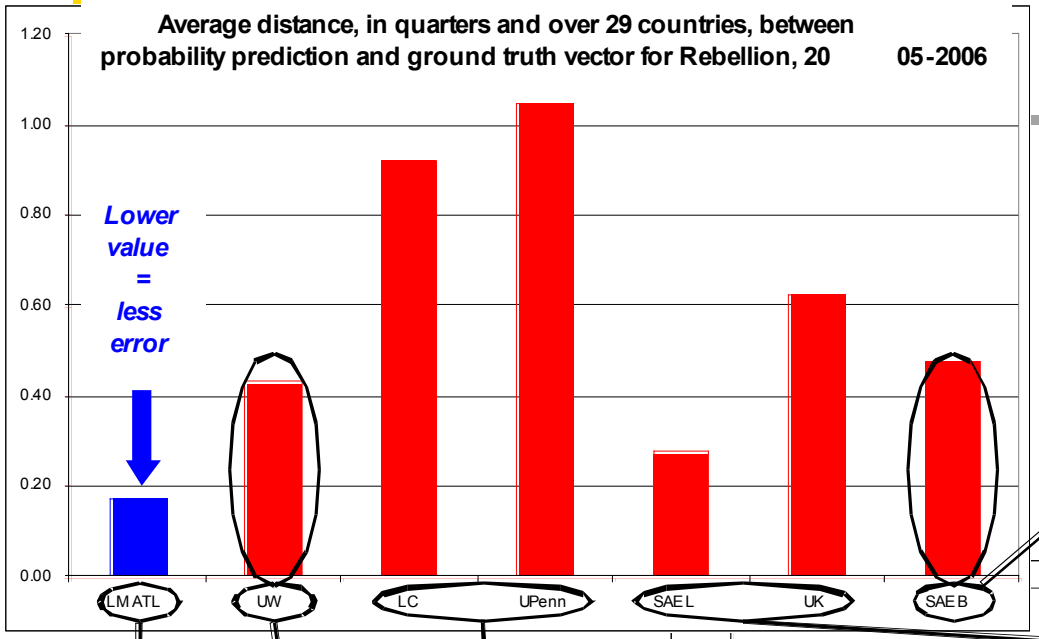
- $\text{Accuracy} = \dfrac{\#\text{ of correct predictions}}{\#\text{ of predictions made}}$
- $\text{Recall} = \dfrac{\#\text{ of correctly predicted conflicts}}{\#\text{ of conflicts that occurred}}$
- $\text{Precision} = \dfrac{\#\text{ of correctly predicted conflicts}}{\#\text{ of conflicts predicted to occur}}$

Phase 1 Results: LM-ATL Out-of-Sample Results (DARPA Chart)



- Exceeds metrics for the maximum intensity index and 3 instability events: Rebellion, Insurgency, and Ethnic/Religious Violence – Passes Phase 1 gates
- By integrating improved versions of best of breed models from multiple perspectives, team achieves more accurate, precise forecasts than any one model alone

LM ATL Results by Model



- **Aggregating model scores with a learned Bayesian network outperforms any one other model**
  - Because different models cover the EOIs and countries with varying levels of performance

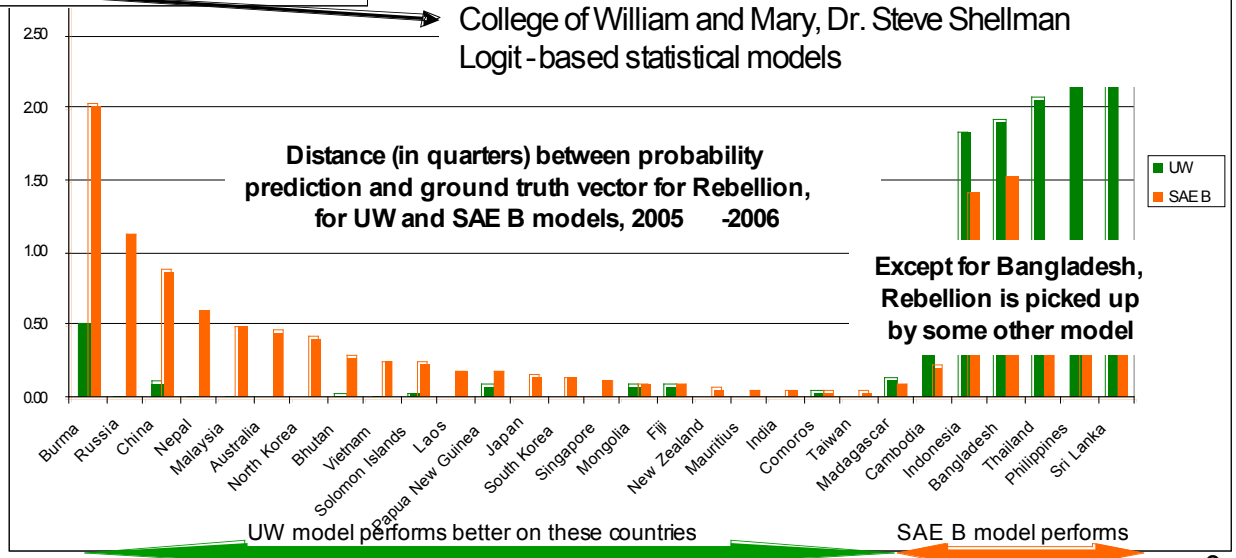
College of William and Mary, Dr. Steve Shellman
Bayesian statistical model

University of Kansas, Dr. Phil Schrodt
College of William and Mary, Dr. Steve Shellman
Logit-based statistical models

Aggregation model

University of Penn, Dr. Ian Lustic and Dr. Barry Silverman
Agent-base models (only 6 and 4 "hard" countries resp.)

University of Washington, Dr. Michael Ward
Geo-spatial statistics applied to trade ties, flow of people, social similarity



Political Instability Task Force (AJPS 2010)

TABLE 2 Out-of-Sample Prediction Exercise for Observed Onsets of Instability, 1995–2004

A. Countries That Had Instability Onsets, 1995–2004. Quintile/decile in model score rankings based on 2-yr. prior data

Year	Top Decile	Second Decile	Second Quintile	Third Quintile
1995	Armenia, Comoros	Belarus		
1996	Albania, Niger, Zambia		Nepal	
1997	Cambodia, Congo-Brazz.			
1998	Guinea-Bissau, Lesotho			Serbia/Montenegro
1999	Ethiopia, Haiti			
2000		Solomon Ils., Guinea*		
2002	Cote d'Ivoire			
2003	Central African Republic			
2004	Iran*	Yemen*		Thailand*

B. Tabulation of All Country-years, 1995–2004. Model estimates based on censored data, using only sample data from prior to year of forecast (countries w/population over 500,000, no ongoing conflict, at least two years old)

	Countries with Instability in $t+2$	Countries Remaining Stable
Predicted for Instability (Top Quintile)	18	233
Predicted for Stability (Not Top Quintile)	3	992
N = 1,246 Percent Classed Correctly	85.7%	81.0%

Number of instability onsets, 1995–2004: 21. Number of instability onsets in top quintile of model scores: 18 (86%).
*Cases added to the problem set in 2005 update.

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 correct	% low correct	% high forecast	% low 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 correct	% low correct	% high forecast	% low 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

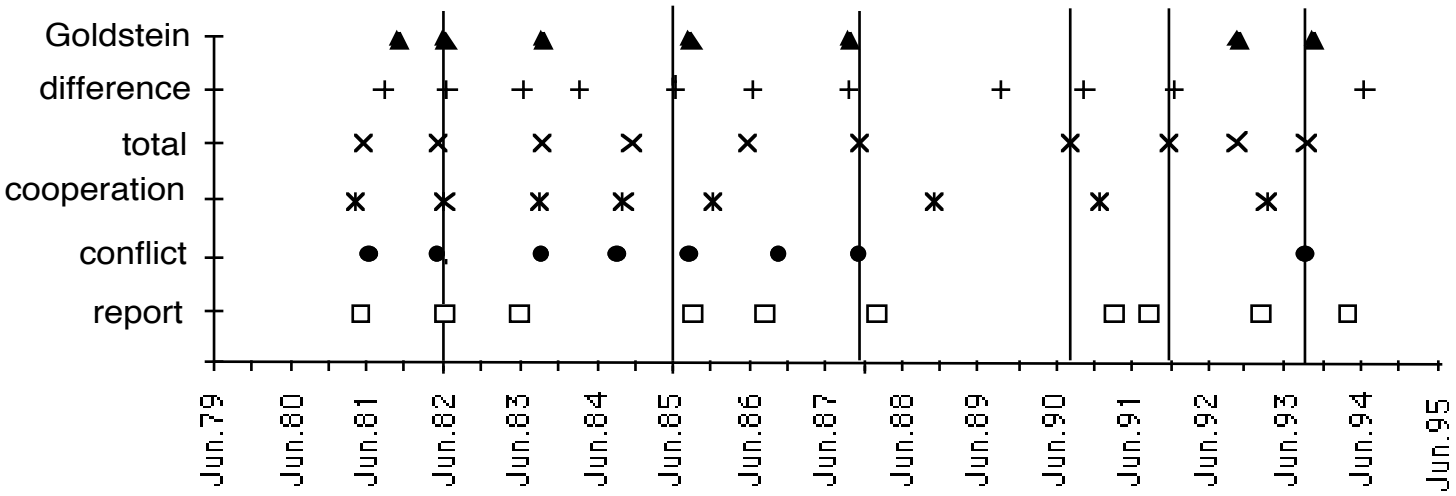
Simplifying Event Scales

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

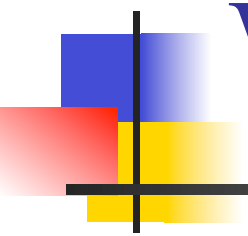
Weighting scheme	%correct	variance explained	canonical correlation	Wilks' λ	significance	# factors
Goldstein	85.6%	76.3%	0.85	0.008	<.001	6
difference	89.7%	74.7%	0.85	0.007	<.001	7
total	94.4%	83.0%	0.93	0.001	<.001	6
conflict	88.2%	76.9%	0.86	0.007	<.001	6
cooperation	92.3%	82.2%	0.91	0.002	<.001	7
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



What’s next?

Newer Statistical Approaches

- Survival analysis (Diehl, Box-Steffensmeier)
  - Predicts likelihood of events as a function of time
- Rare-events analysis (King)
  - Heckman models: two stage process for event occurrence and characteristics of event
  - Zero-inflated Poisson and negative-binomial models
- Non-linear pattern recognition
  - Neural networks (Zeng, PITF)
  - Hidden Markov models (Schrodt, O'Brien)
  - Cluster analysis (Trappl)
  - Reverse Wolfram models (Hudson)
- Social/geographical network analysis (Ward, Gleditsch)
- Bayesian methods (Freeman, Brandt)
  - Uses new data to modify existing assumptions rather than assuming no prior knowledge of the situation

Computational pattern recognition approaches

- Neural networks
- Cluster analysis
  - Support vector machines
  - Correspondence analysis
  - Principal components
  - K-nearest neighbor
  - Latent Dirichlet Allocation models
- Classification and Regression Trees: CART
- Decision trees: ID3/C4.5
- Sequence comparison techniques
  - Hidden Markov models
  - Genetic algorithms

New Directions in Automated Event Coding

- Pre-parsing
- Treatment of actors
- Contextual coding
- High-volume coding
- Server-based integration

Pre-parsing

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

Actor Dictionaries

- mySQL DB contains multiple characteristics of the actor
  - Synonyms ("U.S.", "American"), time-tagged roles, geolocation
- Dictionaries and coders built on the fly, depending of what information will be coded
- Instead of manual, supervised learning on the text, dictionaries developed using
  - Entity identification and coresolution software on the entire text base
  - External lists of actors, e.g. NGOs, rulers.org, CIA World Factbook

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

High-volume, near-real-time coding

- News sources from RSS feeds, news aggregators and other web-based sources
- 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

Questions?

Philip A. Schrodt
Political Science
Pennsylvania State University
State College, PA 16802

Phone: 814-863-8978

Email: schrodt@psu.edu

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