

Technical Political Forecasting using Event Data

Philip A. Schrodt
Department of Political Science
Pennsylvania State University
schrodt@psu.edu
<http://eventdata.psu.edu>

Presentation at the U.S. Holocaust Memorial Museum,
Washington DC
21 September 2012

Main points

- Event data analysis, originally developed under DARPA funding 1965-1980, is a well-understood technique for collecting systematic information on political interactions over time
- Contemporary automated coding methods allow data to be collected in a transparent and reproducible manner in real time at a very low marginal cost
- Statistical models provide 70%-80% accuracy in predicting violence in protracted conflicts in out-of-sample tests at policy-relevant forecast leads
- Exponential increases in the availability of information on political events has produced a major change in the viability and utility of these methods

Integrated Conflict Early Warning System

- Unclassified project funded by DARPA Information Processing Techniques Office
- Funding at \$40-million for 2007-2011
 - Largest quantitative conflict analysis project since the 1970s
- Objective is real-time forecasting of indicators of political instability in Asia with 6-24 month leads, 70%-80% accuracy
- Machine-coded event data has proven to be the core methodology for accurate forecasts
- Data covers 1997-present with 8.5-million stories from 27 sources
- Model accuracy has been assessed with a strict split-sample design

Reference:

Sean O'Brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87-104, 2010.

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

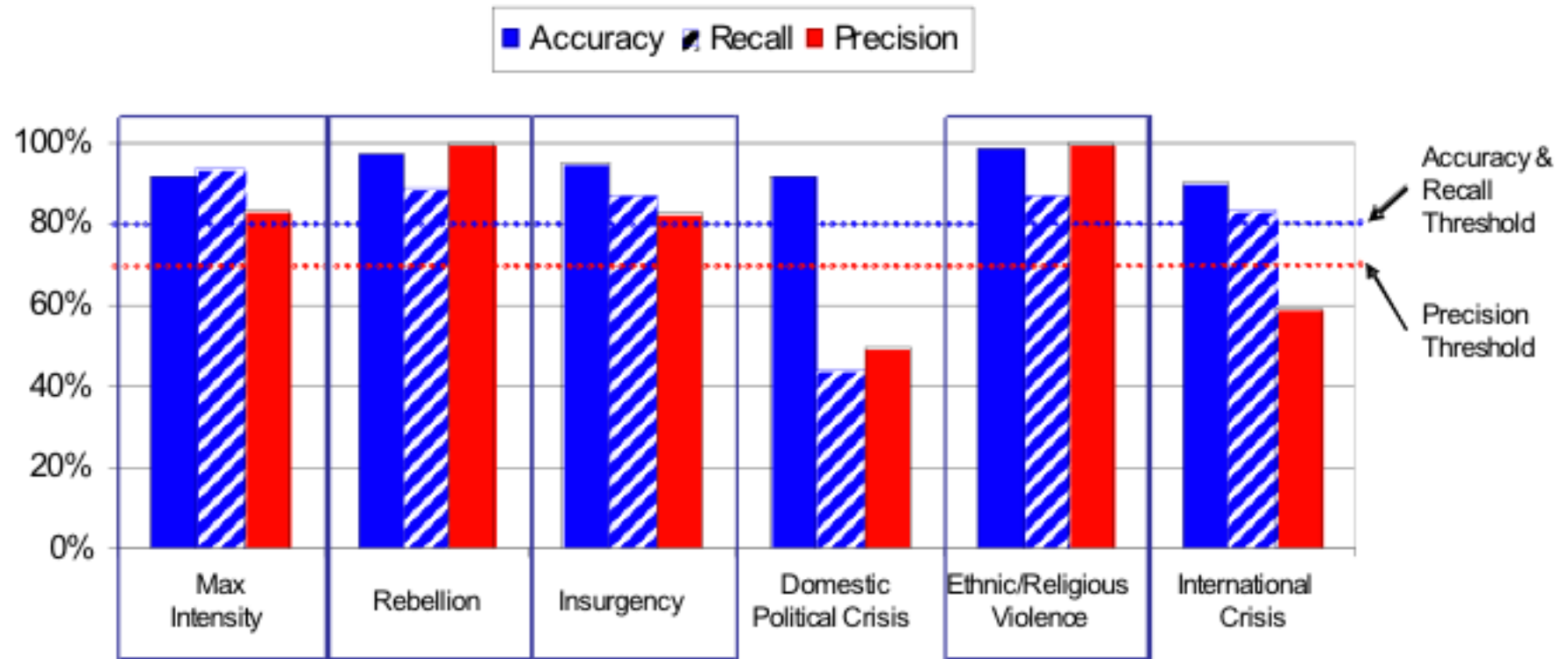
ICEWS Evaluation Criteria

$$\textit{Accuracy} = \frac{\textit{number of correct predictions}}{\textit{total predictions made}}$$

$$\textit{Recall} = \frac{\textit{number of correctly predicted conflicts}}{\textit{total conflicts that occurred}}$$

$$\textit{Precision} = \frac{\textit{number of correctly predicted conflicts}}{\textit{total conflicts predicted}}$$

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



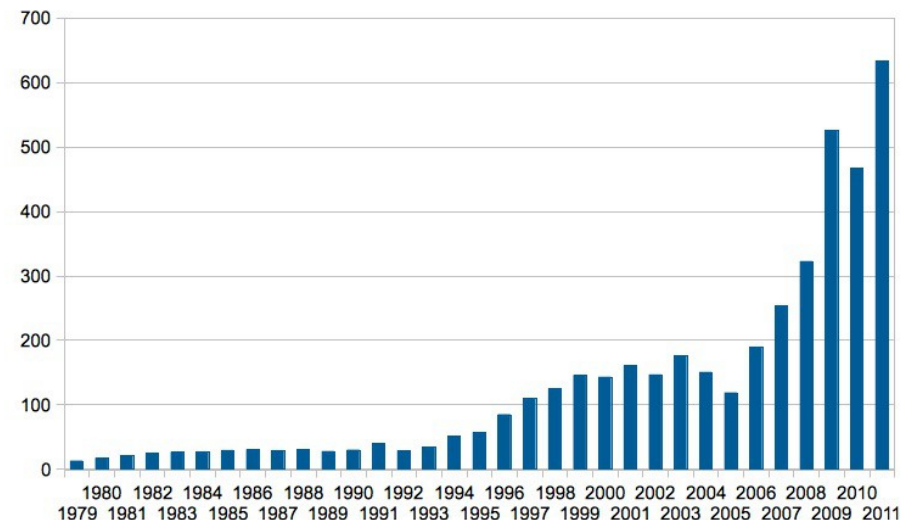
- Exceeds metrics for the maximum intensity index and 3 instability events: Rebellion, Insurgency, and Ethnic/Religious Violence – Passes Phase 1 gates
- By integrating improved versions of best models from multiple perspectives, team achieves more accurate, precise forecasts than any one model alone

ICEWS Phase 1 Event Data

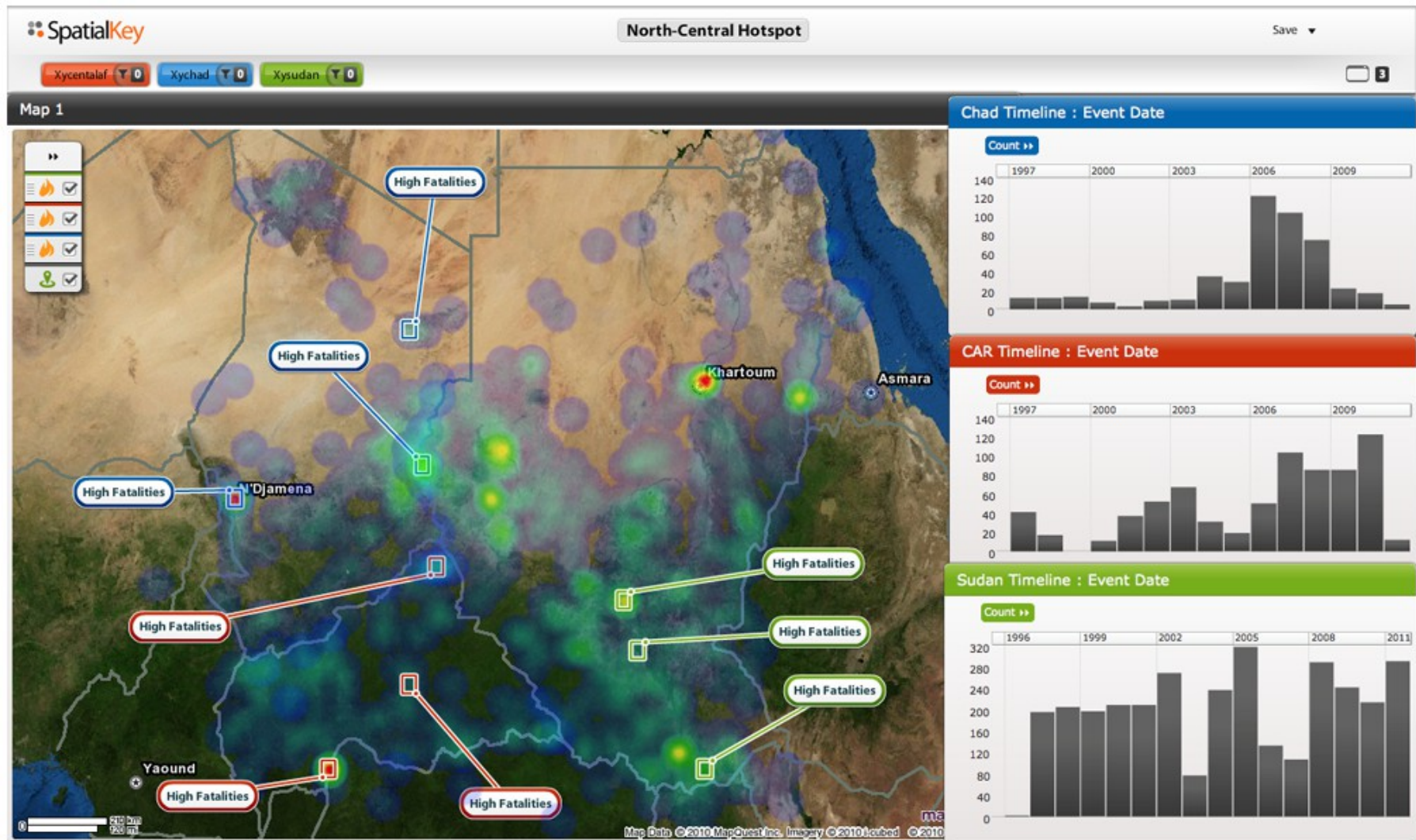
- 30-gigabytes of text from Lexis-Nexis
- 25 sources
- 8-million stories
- 26-million sentences
 - Only first four sentences coded in each story
- 3-million events
- Generally two orders of magnitude greater than any prior event coding effort

GDELT Event Data

- Collaboration of Univ of Illinois supercomputer center (Kalev Leetaru) and Penn State
- 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
- Current intake of 20,000 to 100,000 stories per day
- Planned quarterly updates, backf t to 1800



ACLED: Geospatial



Uppsala Georeferenced Event Dataset

Welcome to the UCDP GED - Uppsala Conflict Data Program's Georeferenced Event Dataset

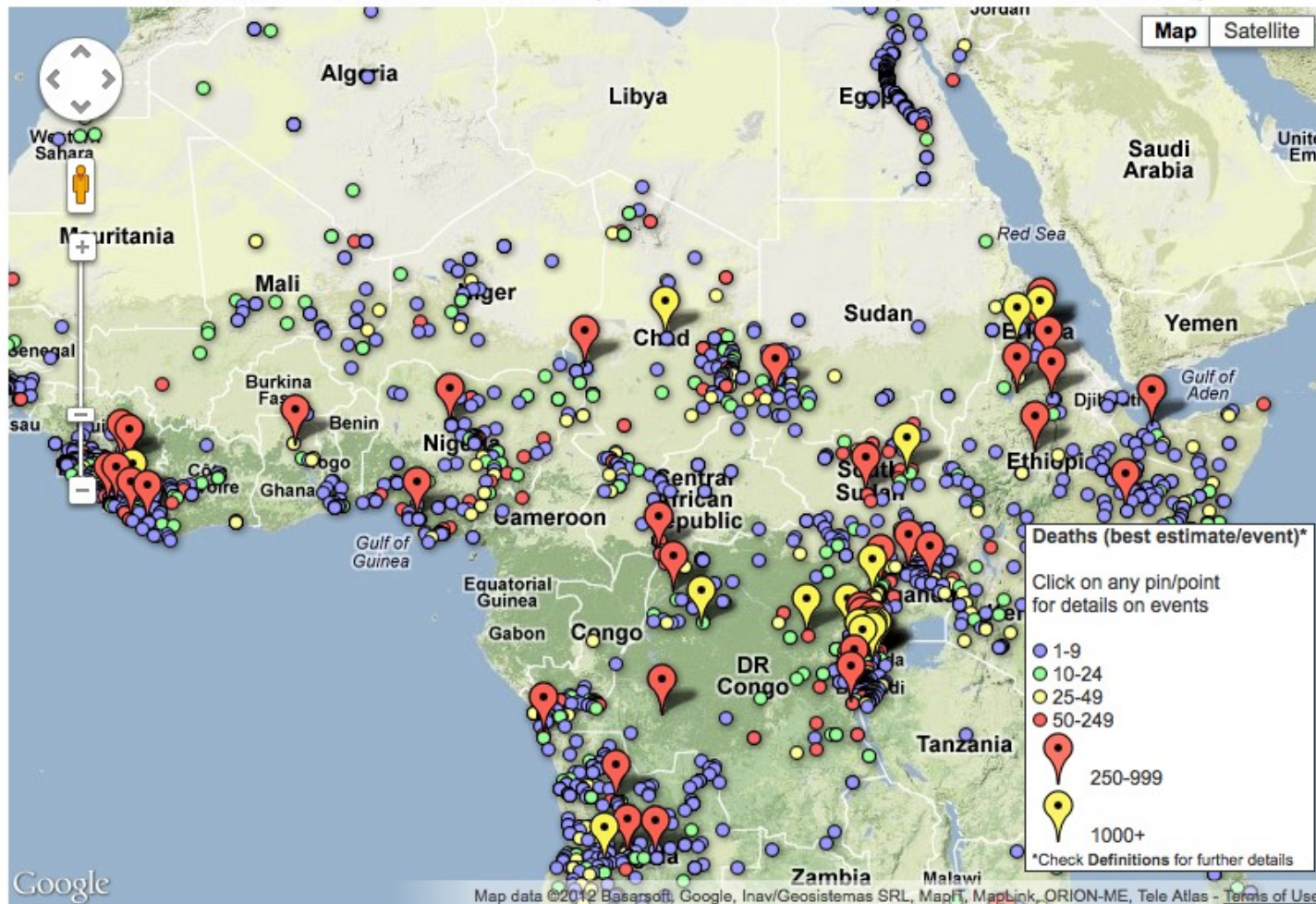
The interactive map below covers all the locations geocoded as part of GED. **Drag and move the red and grey thumbs of the slider below to filter for particular intervals, years, or even particular months or days.** Use the filters to select only one of the three UCDP types of violence. Clicking on each dot or pin will display information regarding events taking place in those respective locations. [Download data here.](#)

1989

2010

Filter for:

☒ All violence ☐ State-based conflicts only ☐ Non-state conflicts only ☐ One-sided conflicts only



The GED is the product of two and a half years of work at the **Department of Peace and Conflict Research, Uppsala University**. The UCDP GED contains conflict data disaggregated spatially and temporally down to the level of the individual incidents of violence. For more details please see the [About UCDP](#) link above.

Political Data Sources

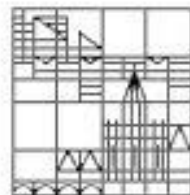


Minorities at Risk



An Open-Source Application for
Publishing, Citing and Discovering Research Data

Universität
Konstanz



Example: 18 December 2007

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

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

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

New York Times, 18 December 2007

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

TABARI Coding

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.

TABARI Coding: Verb

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

Event Code: 111

TABARI Coding: Actors

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

Event Code: 111

Source: IRQ

Target: TUR

TABARI Coding: Agents

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

Event Code: 111

Source: IRQ GOV

Target: TUR

TABARI Coding: Verb [2]

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

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

TABARI Coding: Actors [2]

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

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

Source: TUR

Target: IRQKRD

TABARI Coding: Agents [2]

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

Event Code: 111

Source: IRQ GOV

Target: TUR

Event Code: 223

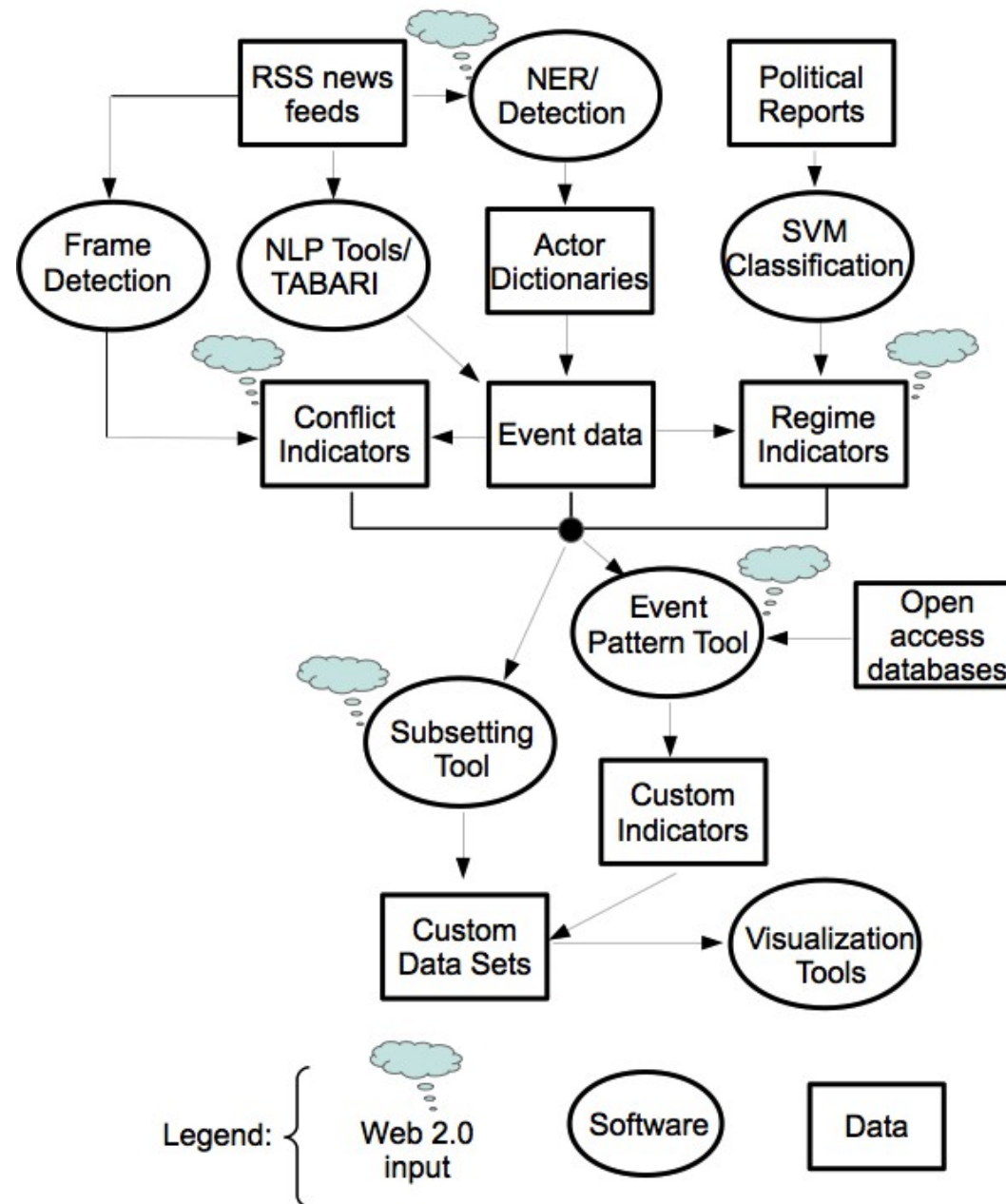
Source: TUR

Target: IRQKRD REB

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



Coding Event Data: Opportunities

- Use of existing political entity lists such as rulers.org, CIA World Factbook, Joshua Project, various non-state actor databases (corporations, IGOs, NGOs, militant groups)
 - GDELT, using only general actor dictionaries and WordNet synonym sets, has a 70% to 80% correlation with KEDS data generated with specialized dictionaries
- Rapid development of remaining actor dictionaries using named-entity recognition tools, possibly with some crowd-sourcing for difficult cases
- GDELT has demonstrated that geocoding is possible using open-source geographical databases
- Incorporation of open-source natural language processing software would improve coding accuracy and consistency

Coding Event Data: Opportunities

- Automated coding can maintain a consistent list of sub-state actors, including date restrictions. This can be modified as one gains experience with the conflict, then the data can be re-coded
- Automated methods are reasonably good at generating numerical totals for deaths and demonstration sizes
 - Duplicated stories are a challenge here
- In many parts of the world, local newspapers are now available electronically, particularly in real-time
 - Coding of non-English sources through machine translation is a possibility
- Thematic coding of political statements is a possibility
- CAMEO coding scheme could be augmented to code “stressors” such as natural disasters and disease, financial instability, and non-political criminal activity



Science

11 February 2011 | \$10

example knowledge

scientific
information
research
data
climate
time
challenges
tools
analysis
visualization
human
become
2010
et
make
AAAS

European Media Monitor



Top Stories

UPDATED EVERY 10 MINUTES, 24 HOURS PER DAY.

Search

Main Menu

Top Stories

24 Hours Overview

Events Detection

Most Active Themes

Mobile Edition

Help about EMM

Overview

Advanced Search

Sources list

Web Site Map

EU Focus

EU Policy Areas

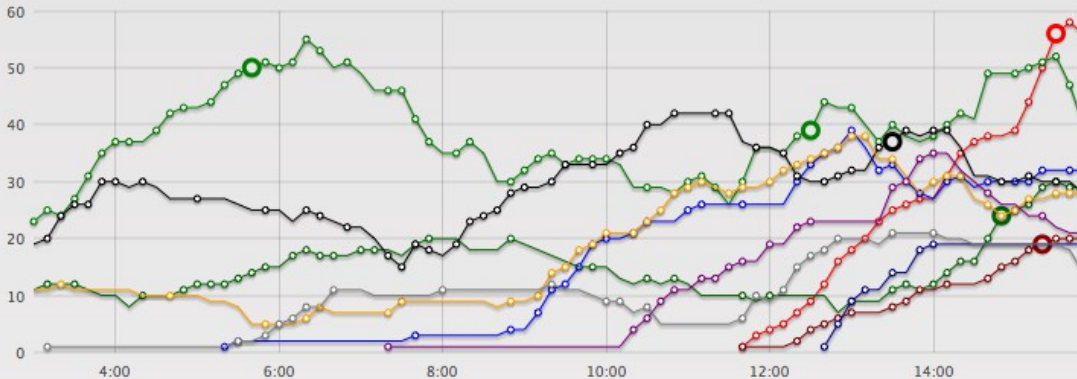
Themes

The World

Offices & Agencies

Current top 10 stories

Language: en Period: Aug 22, 2012 9:50 AM – Aug 22, 2012 9:50 PM



UN chief to visit Iran in defiance of US, Israel

Articles: 59 | Last update: Aug 22, 2012 9:48:00 PM | Start: Aug 22, 2012 5:47:00 PM | Sources: 40 | Peak: 1 | Current rank: 1

UN's Ban to Attend Iran Summit Over Israel's Opposition

bloomberg Wednesday, August 22, 2012 9:48:00 PM CEST | info [other]

Secretary-General Ban Ki-moon will attend a Non-Aligned Movement summit in the Iranian capital Tehran this month in a decision that was made over Israel's opposition and to the chagrin of the U.S. Ban "takes seriously his responsibility and that of the United Nations to pursue diplomatic engagement.....

More articles...

Greece seeks more time; Juncker meets with Samaras

Articles: 175 | Last update: Aug 22, 2012 9:42:00 PM | Start: Aug 22, 2012 12:09:00 AM | Sources: 76 | Peak: 57 | Current rank: 2

Samaras Greece needs air to breathe

Irishsun Wednesday, August 22, 2012 9:42:00 PM CEST | info [other]

Antonis Samaras heads to Berlin on Friday for talks with German Chancellor Angela Merkel, whose country is the largest single contributor to Greece's (EURO)240 billion (\$300 billion) bailout

More articles...

King 'interest in accused athletes'

Articles: 61 | Last update: Aug 22, 2012 9:43:00 PM | Start: Aug 22, 2012 11:28:00 AM | Sources: 39 | Peak: 3 | Current rank: 3

Jordanian Paralympians Are Granted Bail

wsj-en Wednesday, August 22, 2012 9:43:00 PM CEST | info [other]

Two Jordanian Paralympic powerlifters and a trainer were granted bail Wednesday after being accused of sexual assault in Northern Ireland....

More articles...

Tools

Wednesday, August 22, 2012 10:12:00 PM CEST

RSS | KML | MAP

subscribe | manage

info

Languages

Select top stories in other languages.

ar

bg

cs

da

de

el

en

es

et

fi

fr

hr

hu

it

lt

lv

mt

nl

pl

pt

ro

ru

sk

sl

sv

sw

tr

zh

Show additional languages

Interface:

en – English

Legend


Country Watch

The country most in the news at the moment.

Greece (GR)

News about Greece

News sources from Greece



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>

Ideal characteristics technical forecasting models

- Ability to predict at least as well as a human analyst
 - This may be easier than we think. Tetlock estimates the predictive accuracy of human political analysts is only slightly greater than 50% for difficult forecasting problems.
 - Taleb and Kahneman discuss the “narrative fallacy” which makes human forecasts appear to be far more accurate than they actually are
 - Early work (1980s) in “expert systems” for classification problems showed the systems tended to be about 10% more accurate than humans and achieved this accuracy using less information.
- Known probabilities of forecasting outcomes—true positives, false positives, false negatives, true negatives—with confidence intervals
- Consistent and affordable real time updating

Methods of Modeling

- Machine Learning
 - Very flexible methods of determining relationships
 - No [widely accepted] theory of error
- Classical (“frequentist”) statistics
 - Objective is determining whether a variable has a *non-zero* effect: “significance tests”
 - Effective in experimental and other randomized settings, but generally useless for predictive models
- Bayesian statistics
 - Objective is determining a probability of the impact of a variable based on your prior knowledge and the new data
 - Corresponds to how most people actually think about data but has only become computationally feasible in the past twenty years

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, which tend to converge on about 80% accuracy
 - Political Instability Task Force
 - ICEWS
- Decision-makers now expect visual displays of analytical information, which in turn requires systematic measurement
 - “They won’t read things any more”
- “Big Data” analytical methods
 - *Possible* applications to social media, though this is unproved

Contemporary Technical Political Forecasting

- State Failures Project 1994
- 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-present
- Political Instability Task Force 2002-present
- DARPA ICEWS 2007-present
- Peace Research Center Oslo (PRIO) and Uppsala University UCDP political forecasting models

Questions?

Philip A. Schrod

Political Science

Pennsylvania State University

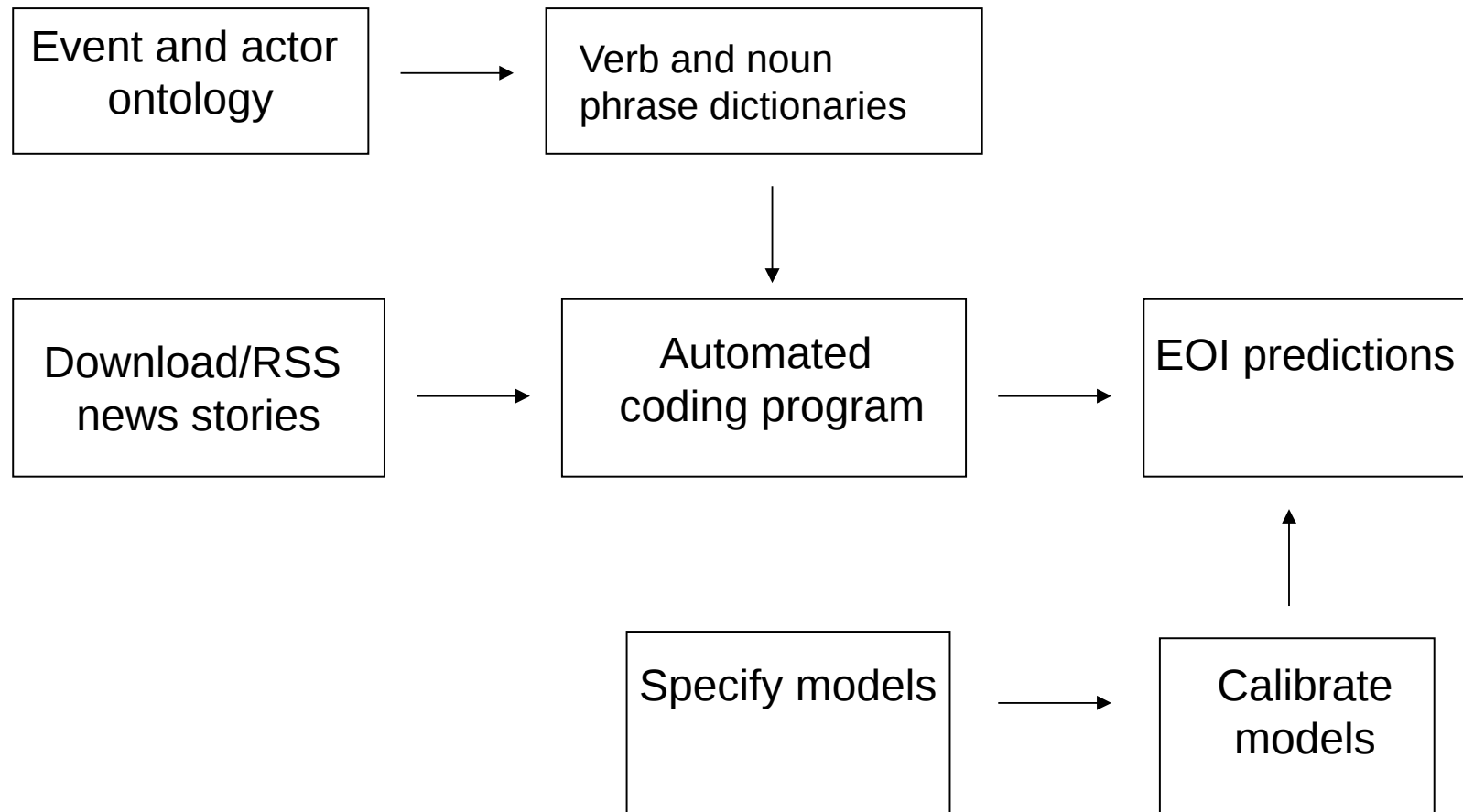
State College, PA 16802

Phone: 814-863-8978

Email: schrod@psu.edu

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

Event Data Generation Process



Getting to real time event data: Updating and calibration

- How frequently do the automated coding dictionaries need to be updated manually?
 - This depends heavily on the success of the automated actor detection software
 - Affective story classification models will also require occasion updating given rhetorical shifts, though some of this can be automated
- How frequently do the models need to be updated?
 - This is a tradeoff between the information available in a large data set versus the currency of using the most recent data

Open sources



Mainstream media



Internet and
new social media

Questions?

Philip A. Schrodtt

Political Science

Pennsylvania State University

State College, PA 16802

Phone: 814-863-8978

Email: schrodtt@psu.edu

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

New Social Media

Police tell high school students to disguise their identity on FaceBook

+

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

OMG! Jihadis are checking out the junior prom!