Open Event Data and the Prospects for Near-Real-Time Forecasting Models

Philip A. Schrodt

Parus Analytics
Charlottesville, Virginia, USA
schrodt735@gmail.com

Paper presented at the Conference on Forecasting and Early Warning of Conflict and Political Instability
Peace Research Institute, Oslo
24 April 2015

Technical Political Forecasting: The Debate

▸ ARGUMENT PRINT | TEXT SIZE | EMAIL | SINGLE PAGE

Why the World Can't Have a Nate Silver

The quants are riding high after Team Data crushed Team Gut in the U.S. election forecasts. But predicting the Electoral College vote is child's play next to some of these hard targets.

BY JAY ULFELDER | NOVEMBER 8, 2012

Vs.

▸ ARGUMENT PRINT | TEXT SIZE | EMAIL | SINGLE PAGE

Predicting the Future Is Easier Than It Looks

Nate Silver was just the beginning. Some of the same statistical techniques used by America's forecaster-in-chief are about to revolutionize world politics.

BY MICHAEL D. WARD , NILS METTERNICH | NOVEMBER 16, 2012

Data!


Science
11 February 2011 | $10
scientific
data
information
climate
research
science
new
analysis
visualization
researchers
access
many
human
knowledge
example
AAAS


PRIO


ACLED
Armed Conflict Location & Event Dataset


Minorities at Risk

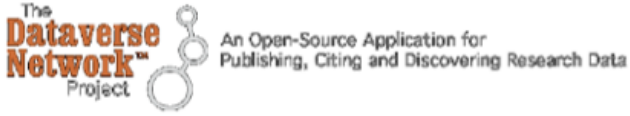

CORRELATES OF
WAR


UPPSALA
UNIVERSITET


PENNSTATE
1855


UCDP
GEOREFERENCED EVENT DATASET


CIRI Human Rights Data Project
www.humanrightsdata.org


The
Dataverse
Network
Project
An Open-Source Application for
Publishing, Citing and Discovering Research Data


Polity
IV


GTD
Global Terrorism Database


Universität
Konstanz

Open Source Software

MySQL

python

mozilla Firefox

ubuntu

source forge

T R

perl

Ruby PROGRAMMING Language

WordPress.com

APACHE HTTP SERVER

R

LibreOffice The Document Foundation

GNU run free run GNU

stackoverflow

FREE SOFTWARE FOUNDATION

This must be important: it's in *The Economist*!

The science of civil war

What makes heroic strife

Computer models that can predict the outbreak and spread of civil conflict are being developed

Apr 21st 2012 | from the print edition

Large Scale Conflict Forecasting Projects

- State Failures Project 1994-2001
- FEWER [Davies and Gurr 1998]
- Center for Army Analysis 2002-2005
- Swiss Peace Foundation FAST 2000-2008
- Political Instability Task Force 2002-present
- DARPA ICEWS 2007-present
- IARPA ACE and OSI 2012-present
- Peace Research Center Oslo (PRIO) and Uppsala University UCDP models
- US Holocaust Memorial Museum Prediction Poll 2015
- EU JRC Global Conflict Risk Index 2014-present

Convergent Results

- Most models require only a small number of variables
- Indirect indicators—famously, infant mortality rate as an indicator of state capacity—are very useful
- ’‘Bad neighborhood” geographical effects are large
- Multiple modeling approaches generally converge to similar accuracy
- Statistical challenge: most interesting events are very rare
- 80% accuracy in the 6 to 24 month forecasting window occurs with remarkable consistency: few if any replicable models exceed this, and models below that level can usually be improved
- Forecast accuracy does not decline very rapidly with increased forecast windows, suggesting long term structural factors rather than short-term “triggers” are dominant. “Trigger models” more generally do poorly except as *post hoc* explanations.

What are event data?

Event data reduce news reports to a standard format that can be used in statistical models:

- Date
- Who initiated the action
- Who was the action directed to
- What was done: this is coded into a standard set of categories
- Location of the incident

Most coding has been done using the major international news services: Reuters, Agence France Press, Associated Press, BBC World Monitor and Xinhua.
Current data sets are expanding this to more local sources, and to languages other than English.

Key observation from contemporary event data

The combination of fully automated coding and the increasing number of reports on the web means that we now have an inexpensive "instrument" for systematically monitoring global political behavior in real time.

Why event data are well suited for predicting political change

- Structural indicators such as GDP, infant mortality, regime type, past or adjacent conflict change too slowly
  - They nonetheless affect the overall probability
- Social media indicators change too quickly for long range forecasts
  - This is also a very new type of data
  - Though it may be possible to use aggregate measures
- Newsworthy events are "just right"
  - As existing models have demonstrated
  - Which is why they are "newsworthy"
  - Structural indicators either are reflected in the patterns of events, or can be additional covariates

Phoenix Data System

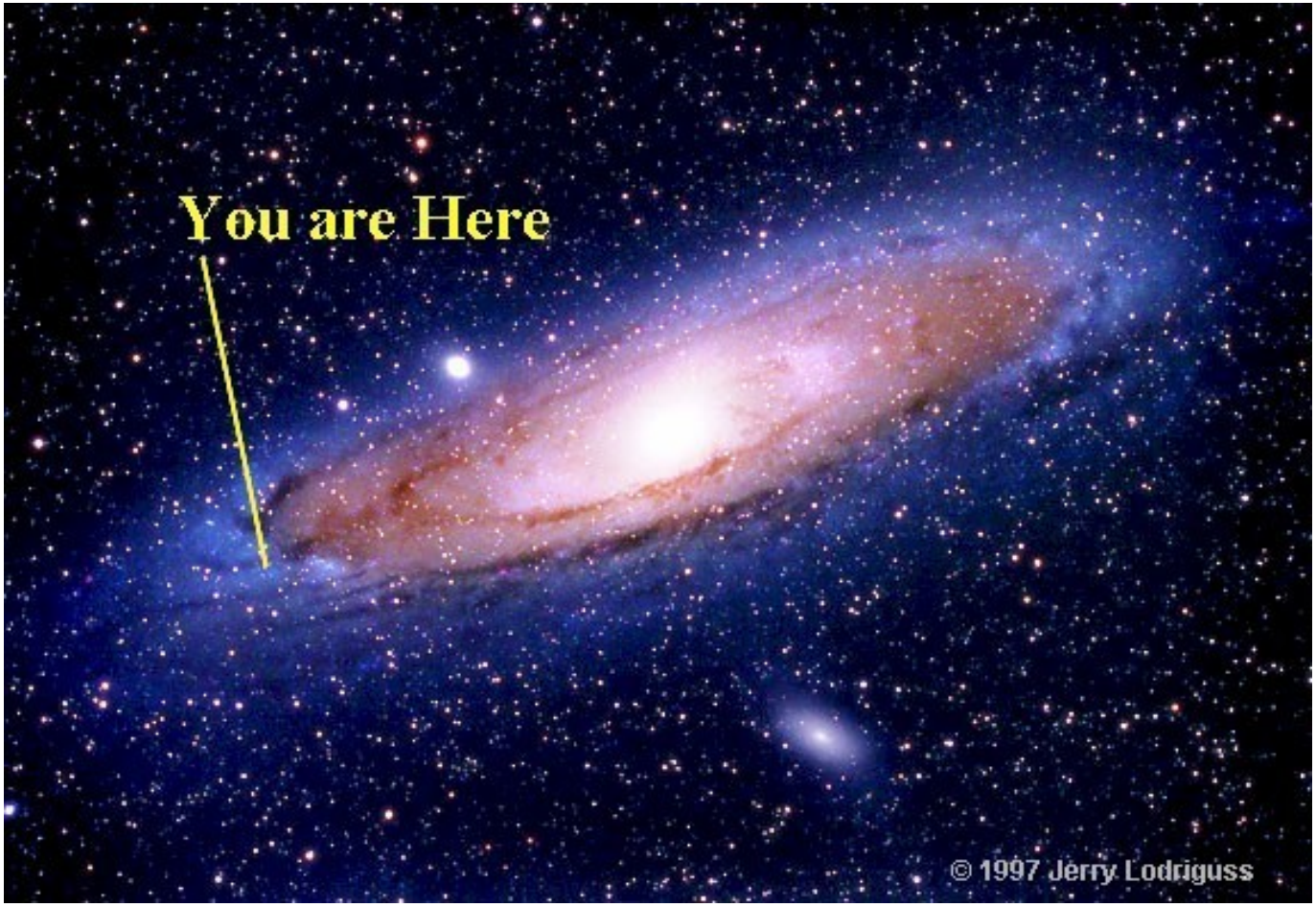
- Open source, open access, open collaboration
- Hosted on GitHub: `https://openeventdata.github.io/`
- Fully modular open-source pipeline to produce daily event data from web sources which can be implemented on inexpensive "cloud" server systems
- Python-language event coder based on the Stanford CoreNLP English-language parser
- Geolocation: "Cliff" open-source system
- Open Event Data Alliance: membership organization to provide at least one source of daily updates with 24/7/365 data reliability. Ideally, multiple such data sets rather than "one data set to rule them all"

ICEWS Data

- U.S. Defense Advanced Research Projects Agency "Integrated Conflict Early Warning System"
- Public version of data released March 2015 covering 1996 to March 2014; monthly updates are anticipated
- 1.6-million events from a variety of sources: `http://thedata.harvard.edu/dvn/dv/icews`
- Included 100,000 entry political actor dictionary
- CAMEO coding system (same as Phoenix, so two data sets should be largely compatible)

Challenge: Black swans

Ideal forecasting targets are neither too common nor too frequent



Final thought:
We will need to learn how to effectively use these tools

Walter Isaacson, *The Innovators*: Throughout the development of computers, there has been a tension between two approaches

- "Artificial intelligence" [Alan Turing, John McCarthy]: Figure out how to get machines to think like humans
- "Computers are tools" [Grace Hopper, Steven Jobs]: Design systems to optimally *complement* human capabilities

Final thought:
We will need to learn how to effectively use these tools

Humans, as social animals, have evolved the ability to construct complex narratives, and in fact derive pleasure from this, as we see from the popularity of stories as diverse as Homer, the *Ramayana* and the Arthurian legands, to the contemporary multi-billion industry in fiction, television and movies.

Computers don't "think" this way—they don't really "think" at all—but for precisely that reason they can give us a different perspective and help us see patterns we might otherwise miss.

Open source approaches allow this approach to be implemented in a variety of different ways, combining the diversity of the human viewpoints with the consistency and transparency of automated approaches.

Thank you

Email:
schrodt735@gmail.com

Slides:
`http://eventdata.parusanalytics.com/presentations.html`

Data: `http://phoenixdata.org`

Software: `https://openeventdata.github.io/`

Papers:
`http://eventdata.parusanalytics.com/papers.html`