POLICY FORUM

POLITICAL SCIENCE

# Growing pains for global monitoring of societal events

## Automated event coding raises promise and concerns

*By* **Wei Wang,[1] Ryan Kennedy,[2] David Lazer,[3,4] Naren Ramakrishnan[1]**

There have been serious efforts over the past 40 years to use newspaper articles to create global-scale databases of events occurring in every corner of the world, to help understand and shape responses to global problems. Although most have been limited by the technology of the time (*1*) [supplementary materials (SM)], two recent groundbreaking projects to provide global, real-time "event data" that take advantage of automated coding from news media have gained widespread recognition: International Crisis Early Warning System (ICEWS), maintained by Lockheed Martin, and Global Data on Events Language and Tone (GDELT), developed and maintained by Kalev Leetaru at Georgetown University (*2*, *3*). The scale of these programs is unprecedented, and their promise has been reflected in the attention they have received from scholars, media, and governments. However, they suffer from major issues with respect to reliability and validity. Opportunities exist to use new methods and to develop an infrastructure that will yield robust and reliable "big data" to study global events—from conflict to ecological change (*3*).

Automated event coding parses individual sentences into SUBJECT VERB OBJECT format and categorizes the action using a framework like CAMEO (Conflict and Mediation Event Observations). So a statement like "Secretary of State John Kerry complained about Russia's support of Syria's Bashar al-Assad" would be coded as US GOVERNMENT/DISAPPROVE/RUSSIAN GOVERNMENT. This can be refined into a numeric level of hostility or cooperation by using scales like the Goldstein Score. Whereas CAMEO focuses on categories for international and domestic conflict, similar frameworks could be developed for almost any kind of interaction in news media (e.g., transactions between businesses or debates over scientific findings).

Uses for the resulting data have been manifold. Hand-coded and automated event data have been used to anticipate conflict escalation (*2*). When combined with statistical and agent-based models, ICEWS claims a forecasting accuracy of 80%. GDELT has been used to track, e.g., wildlife crime and the rise of hate speech following the U.K. Brexit vote.

There are several challenges in the current approach. First, the focus on sentences removes a great deal of context. Event occurrences do not neatly partition into sentences. This lack of context, for example, often fails to distinguish rereporting of historic events, and this results in high rates of duplication.

***Pullquote or liftout quote piece tops on baseline as shown a "synthesis of dummy type goes here."***

Second, event data programs can have inconsistent corpuses over time. For instance, GDELT has expanded the number and variety of its sources. Although expansions are positive—incorporating, for example, more non-Western news sources—there is difficulty interpreting what a spike in GDELT data at a particular time means; the project has not been entirely transparent on how these expansions have taken place. ICEWS has been more consistent about maintaining a common set of sources across nearly 25 years.

Third, the text-processing systems used in event coding are still similar to ones developed more than 20 years ago. Although ICEWS has recently begun leveraging a machine-learning approach, GDELT still relies on dictionary-based pattern matching that leads to overly simplified or misclassified coding instances. The field of text processing has developed a range of tools to address these issues (*4*, *5*). Finally, although there are a few large event-coding programs, the academic groups working on these problems are surprisingly diffuse and isolated (SM).

### RELIABILITY

Our first set of experiments deals with the reliability of event data—whether programs ostensibly using similar coding rules produce similar data. We used four sources of event data [ICEWS, GDELT, Gold Standard Report (GSR), and Social, Political and Economic Event Database (SPEED)], all designed to detect protest events. GDELT and ICEWS are fully automated and are the best attempts so far at real-time global event data. The GSR data set, generated by the nonprofit MITRE Corporation, is hand-coded from local and international newswires in Latin America since 2011 (*6*). SPEED is a semiautomated global event data system by the University of Illinois that uses a combination of human and automated techniques for identifying events. It touts the high validity of its event coding (*7*). GSR and SPEED were developed to provide a "ground truth," but their methods would be difficult and expensive to scale. Although these systems have different origins (e.g., ICEWS was meant to encode strategic interactions, often among nation-states, and GSR was meant to focus on tactical, local issues) (SM), we anticipate that overall there should be a high correlation between the time series of events generated by these projects, even if the event counts are not comparable.

We find that the correlation between event data collections is in the area most consider "weak" [correlation coefficient ($r$) < 0.3]. The average correlation between GDELT and the GSR across Latin America is 0.222, and the correlation between ICEWS and the GSR is 0.229. SPEED and GDELT records match (i.e., both data sets recorded a protest happening on the same day) 17.2% of the time. SPEED and ICEWS agree on 10.3% of events. ICEWS and GDELT rarely agree with each other, producing an average correlation across Latin America of 0.317 (SM). Correlations between countries improve when there are large upticks in event counts. For example, the large uptick in protests in Venezuela in January 2014 is well captured by both ICEWS and GDELT (see the chart). They also improve when the time scales are rougher (from daily to weekly or monthly) (SM). Reliance on English-language news coverage results in stronger correlations for states with more coverage in the Western press (e.g., Brazil) (*8*) (SM).

### VALIDITY

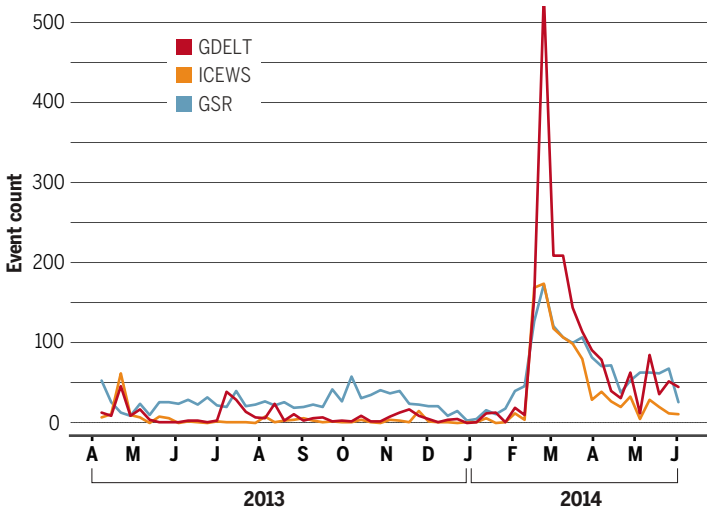
To assess the the degree to which event-coding projects reflect unique real-world events, we leveraged a special characteristic of the GDELT data set. Since its launch in April 2013, GDELT has provided URLs for

[1]*Discovery Analytics Center, Virginia Tech, Arlington, VA 22203, USA.* [2]*Center for International and Comparative Studies, University of Houston, Houston, TX 77204, USA.* [3]*Lazer Lab, Northeastern University, Boston, MA 02115, USA.* [4]*Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. Corresponding author: rkennedy@uh.edu*

most of its coded events. We looked at protest events up to 2 July 2014 (431,549 records), extracted content for records with a valid URL (344,481 records), and filtered them to assess the validity of their classification as protest events. This yielded 113,932 unique, nonduplicated events that are highly probable to be about protests at the time reported.Even for these filtered records, only 49.5% are classified as referring to actual protests, roughly in line with what we found in 1000 human-coded records (SM). After keyword and temporal filtering, de-duplication of events, and machine-learning classification of real events from nonevents or planned events, only 21% of GDELT's valid URLs indicate a true protest event.

## Weekly count of protest events in Venezuela

Event data from GDELT, Global Data on Events Language and Tone; ICEWS, International Crisis Early Warning System; and GSR, Gold Standard Report. (See SM.)




The ICEWS system was more robust (about 80% of keyword-filtered events were classified as protest events) but still vulnerable to duplicate events (<20% of the recorded events). Thus, computer-automated event data often duplicate and misclassify events, and tools, including the ones used here, deal with many of these issues (*4*, *5*). Similar tests for the other 19 event categories in GDELT and ICEWS found similar problems (SM).

## POLICY IMPLICATIONS

Coding interactions in news media is complex, as it involves actor recognition and normalization, time-frame detection, geocoding, event encoding, classification, multilingual support, and other issues. Yet the history of event data has been one of small teams and underfunded research. It has not helped that much of the development has taken place in political science, a discipline under constant threat of having its National Science Foundation (NSF) funding cut by Congress.

As scholars and government agencies create the next generation of large-scale event data, two goals should shape their efforts. First, new efforts must develop a multidisciplinary community of scholars, including computational linguists, data analytics professionals, information extraction practitioners, and domain experts. Although there have been improvements in the natural language processing used for ICEWS (*9*), innovation in the event data field has been slow, especially in handling contextual features and temporal and geographic information. Neither ICEWS nor GDELT were designed to deduplicate events; multiple occurrences have sometimes been used to denote event significance and to support improved modeling. The one sentence per event model is not sufficient for predictive, diagnostic, or abductive reasoning. Research on probabilistic databases can help one reason about inconsistency issues in information extraction and how best to integrate imprecise information into event coding (*10*, *11*). It is time to develop a strong community of teams competing to create the best possible event data, and event-coding software should be released publicly to encourage community engagement.

Second, the corpus used to create event data must be made explicit, and, to the extent possible, shared between teams. As demonstrated by legal issues faced by GDELT [a dispute over use of source materials resulted in scholars abandoning the project and obstacles to using the data for publication (SM)], the current system, where corpus development can only be done by well-funded individual teams with exclusive rights to material, is problematic and encourages atomization of the field. Such a corpus should include more non-English sources to avoid some of the issues observed above (SM).

We recommend developing open test beds of event data against which different approaches can be tested. These test beds should be composed of a representative set of textual data, with some portion hand-coded. Such test beds can be used in contests, like those sponsored by DARPA (Defense Advanced Research Projects Agency) or TREC (Text Retrieval Conference), where different approaches to text analysis compete to produce the best automated coding for event data. This would allow scholars to test tools already developed for text analysis in other areas and to produce tools that deal with tracking interactions from media reports.

A consortium should be developed to provide real-time controlled access to a comprehensive array of copyrighted material, protect the business interests of news agencies, and elicit broader social interest in event data. The UN Global Pulse initiative and Flowminder in Sweden, which address similar issues regarding cell phone data, could provide a model.

Programs like those proposed here have been tried in other areas, such as social media analysis and search-engine technology, with strong results (*12*). Such an effort can go a long way toward settling the debate over the extent to which fully automated approaches, like those of GDELT and ICEWS, can compete with semiautomated approaches like that of SPEED. Event data can provide insights into a range of global problems, from national security to the spread of diseases. Our ability to reason about world affairs would be improved by the availability of high-quality event data. ■

## REFERENCES AND NOTES


1. P. A. Schrodt, D. J. Gerner, *Am. J. Pol. Sci.* **38**, 825 (1994); http://eventdata.psu.edu/papers.dir/automated.html.
2. S. P. O'Brien, *Int. Stud. Rev.* **12**, 87 (2010).
3. K. Leetaru, P. A. Schrodt, paper presented at the International Studies Association Annual Convention, San Francisco, CA, 3 to 6 April 2013 (ISA, Storrs, CT, 2013); http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf.
4. R. Huang, E. Riloff, in *Proceedings of the 26th Conference of the Association for the Advancement of Artificial Intelligence*, Toronto, Ontario, Canada, 22 to 26 July 2012 (AAII, 2012).
5. D. McClosky et al., in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT 11)*, Portland, OR, 19 to 24 June 2011 (Association for Computational Linguistics, Stroudsburg, PA, 2011), pp. 1626–1635.
6. N. Ramakrishnan *et al.*, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, 24 to 27 August 2014 (ACM, 2014), pp. 1799–1808.
7. P. F. Nardulli et al., *Sociol. Methodol.* **45**, 148 (2015).
8. N. B, Weidmann. *J. Conflict Resolut.* **59**, 1129 (2015),
9. E. Boschee et al., in *Handbook of Computational Approaches to Counterterrorism* (Springer, New York, 2013), pp. 51–67.
9. R. Gupta, S. Sarawagi, in *Proceedings VLDB '06*, Seoul, Korea, 12 to 15 September 2006 (Very Large Database Endowment, 2006), pp. 965–976.
10. J. Pujara et al., *AI Mag.* **36**, 65 (2015).
11. See, for example, Intelligence Advanced Research Projects Activity's (IARPA's) Open Source Indicators program, https://www.iarpa.gov/.


## ACKNOWLEDGMENTS


Research was supported in part by the IARPA via Department of the Interior, National Business Center (D12PC000337) and NSF (grant 1125095). Views and conclusions contained herein are those of the authors and should not be interpreted as representing official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, the NSF, or the U.S. government.


## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/issue/page/suppl/DC1


10.1126/science.aaf6758


GRAPHIC: N. CARY/*SCIENCE*