Managing Big Data For Results

Story
Slides

New Feature in Spotfire 5.5
Add TIBCO Spotfire Maps
Browse
Use in Map Chart Properties

Cover Page GDELT 2012 "Reduced" Data Set
GDELT 1979 "Reduced" Data Set
GDELT Data Ecosystem
GDELT Published Paper Tables
GDELT Code Tables
Cover Page GDELT 2011 "Reduced" Data Set
Cover Page GDELT Selected Complete Data Sets
GDELT Year 2001 Complete
GDELT Year 2012 Month March Complete
GDELT Day July 2, 2013 Complete

World Countries

World Cities

Spotfire Dashboards

GDELT

Spotfire Dashboards

GDELT2011

Spotfire Dashboards

GDELTSelected

Research Notes

Meetup Message Board Update

About GDELT: The Global Database of Events, Language, and Tone

What is GDELT?
How to cite GDELT

Data Sources
2.0.0 Introduction

Figure 1: Distribution of GDELT events over time, Mb per year

2.0.1 Agence France Presse

Figure 2: Articles per month Agence France Presse
Table 1: Top 25 countries by percent of all Agence France Presse articles mentioning that country

2.0.2 Associated Press

Figure 3: Articles per month Associated Press
Table 2: Top 25 countries by percent of all Associated Press articles mentioning that country

2.0.3 Xinhua

Figure 4: Articles per month Xinhua
Table 3: Top 25 countries by percent of all Xinhua articles mentioning that country

2.0.4 Comparing the Sources

Figure 5: Z-scored articles per month comparing Agence France Presse, Associated Press, and Xinhua
Figure 6: Combined articles per month across Agence France Presse, Associated Press, and Xinhua

2.0.5 Additional Post Filtering

Figure 7: Agence France Presse events per month by Quad Class
Figure 8: Associated Press events per month by Quad Class
Figure 9: Xinhua events per month by Quad Class
Table 4: Breakdown by percent of all events in each newswire in each Quad Class

3 Automated Coding Engines

3.1 Event Coding

3.1.1 TABARI
3.1.2 Dictionaries
3.2 Location and Tone
3.3 "The Pipeline"
    3.3.1 Pre-processing
    3.3.2 Coding
3.4 One-A-Day Filtering

4 Visualizations
    4.1 Quad Count Time Series for Selected Dyads
    Table 5: Quad Count Time Series for Selected Dyads
    4.2 District-level Conflict in Afghanistan
        Figure 10: The Number of Material Conflict events per Afghani District from 2001 to 2012 (Yonamine 2013)
    4.3 Comparison of GDELT and NGO-based data for City-level violence in Syria
        Figure 11: The Number of GDELT-derived Violent Events in Homs and Aleppo from January 2012 through June 2012 (Yonamine 2013)
        Figure 12: The Number of NGO-reported Violent Events in Homs and Aleppo from January 2012 through June 2012 (Yonamine 2013)
    4.4 GDELT: 29 January 2011
        Figure 13: GDELT for 29 January 2011 based on number of events and Goldstein scores
        Figure 14: GDELT map for 29 January 2011 based on number of mentions of event and Goldstein scores
        Figure 15: GDELT network map for 29 January 2011 based on dyadic links and average Goldstein scores

5 Comparison to Other Data Sets
    5.1 KEDS/Reuters Levant
        Figure 16: Comparison of KEDS/Reuters and GDELT Quadcounts for Israel -> Lebanon
        Figure 17: Comparison of KEDS/Reuters and GDELT Quadcounts for Israel -> Palestine
        Table 6: Hypothesized intervals in Levant data
        Table 7: Comparison of KEDS/Reuters and GDELT Quadcounts by Time Period
        Figure 18: Comparison of KEDS/Reuters and GDELT Quadcounts: 2000-2012 is exceptional
    5.2 ICEWS
        Figure 19: Comparison of ICEWS and GDELT Quadcounts: China->Taiwan 1998-2010
        Figure 20: Comparison of ICEWS and GDELT Quadcounts: China->Taiwan 1998-2010 with V ERCP < 30 cases removed
        Figure 21: Comparison of ICEWS and GDELT Quadcounts: India->Pakistan 1998-2010
        Figure 22: Comparison of ICEWS and GDELT Quadcounts: South Korea->North Korea 1998-2010
        Figure 23: Comparison of ICEWS and GDELT Quadcounts: USA->Japan 1998-2010

6 Future Plans and Collaboration Opportunities
7 Appendix: GDELT CODEBOOK Version 1.0
7.1 INTRODUCTION
7.2 BASE ATTRIBUTES
7.3 ACTOR CODE BREAKOUT
7.4 EVENT ACTION ATTRIBUTES
7.5 EVENT GEOGRAPHY

References

Footnotes

1
2
3
4
5
6
7
8
9
10

Downloading GDELT

"Reduced" 1979-2012 Dataset
Historical 1979 - March 2013 Backfile
Daily Updates April 1, 2013 to Present
CAMEO Lookups and SQL Code for Historical and Daily Updates
   CAMEO Actor Code Lookups
   CAMEO Event Code Lookups
   SQL Bulk Loader Code

FAQ

What sources does GDELT rely upon?
Can you send me the source news articles behind your event database?
How does GDELT compare with the US DOD ICEWS?
How does GDELT compare with other event databases?
I've previously used TABARI for a coding project and have an event database, how can I contribute?
I have some ideas for additional event categories you should add.
I'm a computer science researcher with some NLP tools that might be useful.

Managing Big Data For Results

Big Data: A Revolution That Will Transform How We Live, Work, and Think
   Book Description
   Editorial Reviews
   Reviews
Most Helpful Customer Reviews

"Big Data" is not a new concept...

Agenda

Speakers

Government Speakers
  C. Scot Atkins
  Niall Brennan
  Jack Collins, Ph.D.
  Suzi Iacono, Ph.D.
  Marion Royal

Industry Speakers
  Chid Apte, Ph.D.
  Robert Ames
  Chip Copper, Ph.D.
  Dave Denson
  Bob Kimball
  Steve LeSueur
  Gary Newgaard
  Clark Richey
  Patrick Sims

Presentations

Story

Managing Big Data For Results featured a new book that was not presented by its authors, three government speakers, a technology showcase by "big data" vendors, and a panel of speakers that I had heard before and decided to pass on hearing.

I suggested that Dr. Jack Collins build a data ecosystem (by crowd sourcing) like I have started for SEER and that Niall Brennan look at The Dartmouth Atlas of Health Care results for Medicare data that seem to answer the questions he raises about why hospital admissions are declining.

A data ecosystem consists of three things: making it easy to get the data and documentation (SEER was somewhat that), making it easy to reproduce the results (SEER was easy to do that), and making it easy to see results of others with the data set (SEER was difficult to find that), and then doing that over multiple data sets to see which can be integrated.

One of the panel speakers, Marion Royal, Data.gov Program Manager, told me that OMB was pushing for use of Word Press for government Web sites like Data.gov etc. because it was cheaper and more Wiki-like. This is essentially what I have used for years now in my Web site! For example, I have Data.gov and many other government web sites in my wiki (called MindTouch).
Rutrell Yashin, Contributing Editor with FCW said and asked the following of me:

In last panel discussion at the FCW Executive Briefing on Managing Big Data there was a question about the NSA’s collecting of information from Verizon customers’ telephone calls. Marion Royal and Suzi Iacono both were glad to see that the news brought about a better understanding of metadata – that in fact, the actual content of the phone calls was not collected, rather the metadata. Royal said the folks at Data.gov struggled with defining metadata when they first started Data.gov. Suzi Iacono noted there was a clear understanding it is not the content of the phone being analyzed but the information about the location of the phone call. Yet, I’ve read some stories where experts on metadata say there is a lot that can be collected about a person through metadata of phone numbers, call time and duration, and information about device interactions with cellular towers. Iacono went on to say that there are a lot of new tools that look at metadata and can help analyze, for example heterogeneous data sets and you can discern where is the needle in the haystack in a data set. She said that this can be used for unclassified research such as in climate change, energy, public safety so you don’t have to federate databases or archives. So it appears that metadata can be used for good or for more detailed tracking of individuals, which is ok if the focus is on tracking and stopping the bad guys. What is your take on it and are you familiar with the types of tools Iacono mentioned?

My reply was as follows: Well metadata is ‘data about the data’, and my opinion is that the “NSA’s metadata” is essentially the data about the phone calls, except for the actual conversation text, because they have to use enough of the data to perform a “network analysis” to see which conversations they actually need to listen to. So I felt that NSA’s use of the term metadata was misleading to the public into thinking it was less than it really is.

I think Iacona is referring to the use of a “data ecosystem” that I suggested Dr. Jack Collins do where one gets all of the data elements in all the data bases of interest to see which data sets have data elements in common so they can be integrated. I use a tool called Spotfire to build a data ecosystem and Iacona may having being referring to that since I gave her Big Data Senior Steering Work Group a presentation/demonstration on that this past January.

So the idea is the same here, both the NSA and NSF (science community) need to look at enough of the data, but may call it metadata, to see how to prioritize the big data for analysis of usually a smaller subset of the data.

Originally metadata used to be essentially the minimum information one could/should put online to catalog the data (like a library card catalog) so one to find a data set of interest, which was usually not online itself (like a book on the shelf in the library). Then the original metadata and data were put online, and now the Open Government Data Initiatives are all about having both online so they are easy to find and use, unless they are classified.

Thank you for your help. I have put the finishing touches on my story. Before I file though I just wanted to make sure I understand what you mean by “So I felt that NSA’s use of the term metadata was misleading to the public into thinking it was less than it really is.” Is it that the NSA is downplaying the significance of metadata, because if they do use enough of it and do the network analysis they can actually glean more information about a person from the phone data they have on numbers, time and length of the call (which is what I understand the metadata consists of)? Or is it more NSA use of the word metadata?

More the latter - the NSA says we are only collecting the metadata and not the data - actual phone conversation - but the metadata is not like original metadata - it is actually most of the data about the phone call to see if they need to listen to the actual conversations - the library catalog type metadata is just the book classification and location.

Knowing the call exists, when it was placed and for how long, etc provides more than library catalogues and supports lots of data analyses to gain more data and insights into us than we can get from library catalogs.
Thanks, Brand. You always bring pertinent insight to the subjects I’m writing about.

You are welcome, Rutrell. Instead of NSA calling it metadata. They should call it the “reduced” data set like the GDELT program does for truth in labeling.

The slides from the FCW Executive Briefing Series: Managing Big Data For Results are quite limited.

An example of what I am talking about is GDELT. The Global Database of Events, Language, and Tone (GDELT) is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world over the last two centuries down to the city level globally, to make all of this data freely available for open research, and to provide daily updates to create the first "realtime social sciences earth observatory." Nearly a quarter-billion georeferenced events capture global behavior in more than 300 categories covering 1979 to present with daily updates.

Updates on the status of these systems can be found at http://eventdata.psu.edu/data.dir/GDELT.html, which also contains a "reduced" version of the data set that has only the core event, actor and location variables (but for all countries and years).

A reduced version of the data set which contains the basic event, actor and geolocation variables for all countries for Jan-1979 to June-2012 can be downloaded from a link at http://eventdata.psu.edu/data.dir/GDELT.html. The file is about 650Mb compressed and includes Python programs for doing basic subsetting and generating counts and R graphics: the file \GDELT.reduced.documentation.txt" describes the file format and utility programs; files are in Unix format.

The Big Data Download is a daily digital program that taps into the big data revolution to deliver valuable information to everyday investors and business decision makers. The show draws upon sources including Yahoo! Search data, social sentiment and investing algorithms created exclusively for this program. The example on May ??, 2013 was GDELT.

The Back Files and Daily Updates are available on the new server so I have tried to reproduce and improve on their results by using the following examples:

- GDELT 2012 "Reduced" Data Set (1)
- GDELT 1979 "Reduced" Data Set (1)
- GDELT Data Ecosystem (1)
- GDELT Published Paper Tables (1)
- GDELT Code Tables (1)
- GDELT 2011 "Reduced" Data Set (2)
- GDELT Selected Complete Data Sets (3)
- Annual: GDELT Year 2001 Complete (3)
- Month: GDELT Year 2012 Month March Complete (3)
- Day: GDELT Day July 2, 2013 Complete (3)
- New Spotfire Maps Feature in Spotfire 5.5 for World Countries and Cities and ISO3166 Codes (3)

The results are shown below in the Slides of Spotfire screen captures and in live Spotfire Dashboards themselves.

The Spotifre files sizes are as follows:

http://semanticommunity.info/AOL_Government/Managing_Big_Data_For_Results
Updated: Wed, 23 Sep 2015 07:38:19 GMT
Powered by mindtouch
GDELT Day July 2, 2013 Complete is confusing to me because it contains data for other days and years besides July 2, 2013.

The New Spotfire Maps Feature in Spotfire 5.5 for World Countries and Cities and ISO3166 Codes made it easy to visualize the GDELT data sets on the World Country Boundary maps and Spotfire provides Bar Charts and Scatter Plots with Filters for Facteded Search.

I wanted to use the Spotfire Network Analytics but it does not appear on the Toolbar of Spotfire 5.5.

Slides

New Feature in Spotfire 5.5

Add TIBCO Spotfire Maps
Browse

Use in Map Chart Properties
World Countries

World Cities

Spotfire Dashboards

GDELT

For Internet Explorer Users and Those Wanting Full Screen Display Use: Web Player Get Spotfire for iPad App
Meetup Message Board Update

Source: Email, June 22, 2013

You asked that we send you a summary of Meetup message board activity. The following is a list of discussions with new messages posted since June 8, 2013 3:19 AM.

DATA SCIENCE DC

Forum: Data Science DC Discussion Forum

New 1/4 billion global geocoded event dataset (GDELT)


Latest message by James Conkling on June 21, 2013 at 11:32 AM
What is GDELT?

The Global Database of Events, Language, and Tone (GDELT) is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world over the last two centuries down to the city level globally, to make all of this data freely available for open research, and to provide daily updates to create the first “realtime social sciences earth observatory.” Nearly a quarter-billion georeferenced events capture global behavior in more than 300 categories covering 1979 to present with daily updates.

GDELT is designed to help support new theories and descriptive understandings of the behaviors and driving forces of global-scale social systems from the micro-level of the individual through the macro-level of the entire planet by offering realtime synthesis of global societal-scale behavior into a rich quantitative database allowing realtime monitoring and analytical exploration of those trends.

GDELT’s evolving ability to capture ethnic, religious, and other social and cultural group relationships will offer profoundly new insights into the interplay of those groups over time, offering a rich new platform for understanding patterns of social evolution, while the data’s realtime nature will expand current understanding of social systems beyond static snapshots towards theories that incorporate the nonlinear behavior and feedback effects that define human interaction and greatly enrich fragility indexes, early warning systems, and forecasting efforts.

GDELT’s goal is to help uncover previously-obscured spatial, temporal, and perceptual evolutionary trends through new forms of analysis of the vast textual repositories that capture global societal activity, from news and social media archives to knowledge repositories.

Key Features

- Covers all countries globally
- Covers a quarter-century: 1979 to present
- Daily updates every day, 365 days a year
- Based on cross-section of all major international, national, regional, local, and hyper-local news sources, both print and broadcast, from nearly every corner of the globe, in both English and vernacular
- 58 fields capture all available detail about event and actors
- Ten fields capture significant detail about each actor, including role and type
- All records georeferenced to the city or landmark as recorded in the article
- Sophisticated geographic pipeline disambiguates and affiliates geography with actors
- Separate geographic information for location of event and for both actors, including GNS and GNIS identifiers
- All records include ethnic and religious affiliation of both actors as provided in the text
- Even captures ambiguous events in conflict zones (“unidentified gunmen stormed the mosque and killed 20 civilians”)
- Specialized filtering and linguistic rewriting filters considerably enhance TABARI’s accuracy
• Wide array of media and emotion-based "importance" indicators for each event
• Nearly a quarter-billion event records
• 100% open, unclassified, and available for unlimited use and redistribution

How to cite GDELT

To cite GDELT, please cite the 2013 International Studies Association (ISA) announcing the dataset:


Data Sources

Sources that were examined to identify events include all international news coverage from AfricaNews, Agence France Presse, Associated Press Online, Associated Press Worldstream, BBC Monitoring, Christian Science Monitor, Facts on File, Foreign Broadcast Information Service, United Press International, and the Washington Post.

Additional sources examined include all national and international news coverage from the New York Times, all international and major US national stories from the Associated Press, and all national and international news from Google News with the exception of sports, entertainment, and strictly economic news.

Events are actively drawn from local, regional, national, and international mainstream news media outlets from throughout the world, including local domestic sources in almost every country on Earth. We are also actively experimenting with incorporating realtime and hyperlocal social media sources along several dimensions, including understanding the geography of social media and engaging in discussions with citizen crisis mapping organizations. We would love to hear from you.

Who made GDELT?

The GDELT team currently consists of Kalev Leetaru of the University of Illinois, Philip Schrodt of Penn State (PSU), and Patrick Brandt of the University of Texas at Dallas (UTD).

GDELT Terms of Use

GDELT is available for unlimited and unrestricted use for any academic, commercial, or governmental use of any kind without fee. You may also redistribute and republish the data in any form. However, any use or redistribution of the data must include a citation to GDELT and a link to this website.

The GDELT dataset is generated completely by fully automatic software algorithms operating with no human oversight or intervention and is based on global news media reporting. No warranties or guarantees of any kind, express or implied, are offered regarding the accuracy or completeness of the data.
Acknowledgements

We would like to specifically acknowledge the following organizations in making this research possible: BBC Monitoring, Reed Elsevier's LexisNexis Group, Google and Google News, and the School of Economic, Political and Policy Sciences at the University of Texas, Dallas.

Philip Schrodt's contributions to the project were funded in part by National Science Foundation grant SES-1004414 and by a Fulbright-Hays Research Fellowship for work at the Peace Research Institute, Oslo (http://www.prio.no). Patrick Brandt's contributions were funded in part by National Science Foundation grant SES-092105.

All Features

Source: http://gdelt.utdallas.edu/allfeatures.html

Below you will find a complete list of all of the features that appear on the front page.

New Scientist Magazine Maps the Syrian Civil War

Peter Aldhous of New Scientist Magazine put together this impressive interactive map of the Syrian civil war that breaks the violence down by quarter to plot how the conflict has been evolving. This map is an interactive companion to an article by Douglas Heaven in the May 13, 2013 issue of the magazine titled "World's largest events database could predict conflict". The map demonstrates using hexagonal binning to cluster event records due to the high density and tight clustering of events around key cities. You can use the time box in the upper-right of the display to interactively move through the conflict quarter by quarter from 2011 to 2013 to see how the conflict has evolved spatially.
Jay Yonamine of Penn State used GDELT in his doctoral dissertation to explore how the spatial resolution of GDELT can be used to better estimate future violence at the district level in Afghanistan. This sequence of maps created by Joshua Stevens is part of a broader display showing the progression of unrest in the country since the US invasion. In this chapter of his dissertation, Jay explores how the high spatial resolution of GDELT can be used to increase the accuracy of violence forecasts in the country.
Daily Top 2000 Google Earth KML Map

Each morning when we post the latest Daily Update file with the previous day’s events, we rank all of the events by the amount of media coverage each got and then compile a list of the top 2000 events that received the most media attention and put them on a Google Earth KML map. You can click on each point and see a list of events found at that location, along with a link to the news article each event came from. This can be a good way to quickly skim the previous day's events and since it is in the industry-standard Google Earth KML format, you can easily mash it up with your own dashboards and mapping displays!
Rolf Fredheim of Cambridge University created this visually stunning series of network displays of Russia's interactions with the world over the past quarter-century, as viewed through the eyes of GDELT. He even includes a YouTube video showing how the network changes week by week over the entire 1979 to 2012 period. He also discusses his observations of using GDELT for macro-level pattern detection.
David Masad of George Mason University put together this great tutorial demonstrating how to replicate some of Rolf Fredheim's R-based visualizations for those more familiar or comfortable with Python. He creates some really impressive publication-ready visuals and includes all of the source code used to create them, along with copious documentation and discussion, making it easy to get started mapping GDELT.
Alex Hanna at the University of Wisconsin-Madison made this great timeline showing the 2011 Egyptian revolution and the various actor classes involved, as captured through the eyes of GDELT. He discusses using the data to understand social movements and political sociology, especially protest behavior, and leveraging the actor codes to lend detail to the underlying social dynamics.
Rolf Fredheim of Cambridge University put together a fantastic assortment of basic analyses and some sophisticated visualizations of GDELT, including a number of publication-friendly maps and graphs. He also shows how to create some fantastic network visualizations of actor interconnections. He includes all of the source code and descriptions of how he generated each image.

GDELT: Global Data on Events, Location and Tone, 1979-2012

Kalev Leetaru, kalev.leetaru5@gmail.com
Philip A. Schrodt, schrodt@psu.edu
Version 1.0 : March 29, 2013

Paper presented at the International Studies Association meetings, San Francisco, April 2013. The authors would like to specifically acknowledge the following organizations in making this research possible: BBC Monitoring, Reed Elsevier's LexisNexis Group, and Google and Google News. We are also indebted to [Dr.] Jay Yonamine for the visualizations of the Syria and Afghanistan included in this paper, and extensive experimentation with GDELT in his recently-defended dissertation. Schrodt's contributions to the project were funded in part by National Science Foundation grant SES-1004414 and by a Fulbright-Hays Research Fellowship for work at the Peace Research Institute, Oslo (http://www.prio.no). Addresses for authors: Leetaru : Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, MC-493, Champaign, IL 61820-6211 USA; Schrodt: Department of Geography, The Pennsylvania State University, 104 Russell Building, University Park, PA 16802 USA.
Schrodt: Department of Political Science, Pennsylvania State University, University Park, PA 16802 USA. A link to the current version of the GDELT data set can be found at http://eventdata.psu.edu/data.dir/GDELT.html

Abstract

GDELT--Global Data on Events, Location and Tone--is a new CAMEO-coded data set containing more than 200-million geolocated events with global coverage for 1979 to the present. The data are based on news reports from a variety of international news sources coded using the Tabari system for events and additional software for location and tone. The data is freely available and we expect to provide daily updates. This paper describes the news sources and some of their characteristics, the various processing steps that are used in generating the data, some comparisons with the KEDS Levants/Reuters and ICEWS/Asia data sets, and some visualizations. We conclude with an outline of planned enhancements to the data in the near future: these include recoding with new WordNet-enhanced dictionaries, the extension of the CAMEO coding to incorporate codes for financial events, disease outbreaks and natural disasters, and the development of an open-source Python-based successor to Tabari which will use parsed input from existing natural language processing tools.

Due to extensive graphics, the .pdf file for this paper is 16Mb, which exceeds the 4Mb limit of the ISA paper server. A copy of the paper can be downloaded from a link at http://eventdata.psu.edu/papers.dir/automated.html (PDF)

1 Introduction

Political event data have had a long presence in the quantitative study of international politics, dating back to the early efforts of Edward Azar's COPDAB [Azar, 1980] and Charles McClelland's WEIS [McClelland, 1976] as well as a variety of more specialized efforts such as Leng's BCOW [Leng, 1987], though these efforts came to a standstill in the 1980s when funding for the large and costly human coding efforts--much of this provided by the U.S. Department of Defense Advanced Research Projects Agency--came to an end. Nonetheless, by the late 1980s, the NSF-funded 1 Data Development in International Relations project [Merritt et al., 1993] had identified event data as the second most common form of data behind the various Correlates of War data sets used in quantitative studies [McGowan et al., 1988]. The 1990s saw the development of two practical automated event data coding systems, the NSF-funded KEDS [Gerner et al., 1994, Schrodt and Gerner, 1994] and the proprietary VRA-Reader (http://vranet.com; King and Lowe 2004) and in the 2000s, the development of two new political event coding taxonomies--CAMEO [Gerner et al., 2009] and IDEA [Bond et al., 2003]--designed for implementation in automated coding systems. By the 2000s, with the decline of inter-state war, most event data studies shifted to the study of internal conflict, with the major project during this period being the $37-million U.S. Defense Advanced Research Projects Agency (DARPA) Integrated Conflict Early Warning System (ICEWS; O'Brien 2010).

These efforts built a substantial foundation for event data. By the mid-2000s, virtually all refereed articles in political science journals used machine-coded, rather than human-coded, event data. However, the overall investment in machine coding technology remained relatively small. This situation changed dramatically with the DARPA-funded Integrated Conflict Early Warning System project, which invested substantial resources in event data development using automated methods. The key dierence between the ICEWS event data coding eorts and those of earlier NSF-funded efforts was the scale. As O'Brien|the ICEWS project director --notes,

. . . the ICEWS performers used input data from a variety of sources. Notably, they collected 6.5 million news stories about countries in the Pacif Command (PACOM) AOR [area of responsibility] for the period 1998-2006. This resulted in a dataset about two orders of magnitude greater than any other with which we are aware. These stories comprise 253 million lines of text and came from over 75 international sources (AP, UPI, and BBC Monitor) as well as regional sources (India Today, Jakarta Post, Pakistan Newswire, and Saigon Times).

While the original objective of ICEWS was conict forecasting in Asia, the dependence of its most successful forecasting models on event data caused the program to morph into the production of a global event data set for 1996 to 2012, coded with the CAMEO
event scheme and a customized sub-state actor scheme. Unfortunately, while the ICEWS project originally suggested that this data would be released for general use, it now appears to have disappeared into the classified world and there are no indications at present that the data will be available for use outside the U.S. government. Suggesting that the ICEWS models and data are proving to be very useful.

Fortunately, the availability of news texts on the web, along with various NSF-funded open source efforts at coding software and dictionary development, means that unlike the situation in the 1980s, it is possible to produce global data without the necessity of large-scale financial support. This paper will describe GDELT—Global Data on Events, Location and Tone—a new CAMEO-coded data set containing more than 200-million geolocated events with global coverage for 1979 to the present. The data are based on news reports from a variety of international news sources coded using the open-source Tabari system for events and additional software for location and tone. The data will be freely available to researchers at an NSF-funded server we are in the process of setting up at the University of Texas/Dallas--this may be operational by the time this paper is presented, and if not, shortly thereafter--and when fully operational, we expect this system will provide daily updates as well as user-friendly subletting capabilities. This paper describes the news text sources and some of their characteristics, the various processing steps that are used in generating the data, some comparisons with the KEDS Levants/Reuters and ICEWS/Asia data sets, some visualizations and concludes with an outline of future developments that we anticipate completing in the next six months.

2 Text Sources

2.0.0 Introduction

Sources that were examined to identify events include all international news coverage from AfricaNews, Agence France Presse, Associated Press Online, Associated Press Worldstream, BBC Monitoring, Christian Science Monitor, Facts on File, Foreign Broadcast Information Service, United Press International, and the Washington Post. Additional sources examined include all national and international news coverage from the New York Times, all international and major US national stories from the Associated Press, and all national and international news from Google News with the exception of sports, entertainment, and strictly economic news.

The approximate distribution of the events over time is shown in Figure 1, which shows the total size of the files by year. Unsurprisingly, given the very substantial changes over the past two decades in both the international news environment and the availability of news on the web, this is anything but constant, and shows a dramatic increase since the beginning of the twenty-first century.

Figure 1: Distribution of GDELT events over time, Mb per year

In this section, we will look at three of the most important sources—AFP, AP and Xinua detail, both in terms of the availability of the data, the focus by country, and the variations in the level of coverage over time.
2.0.1 Agence France Presse

Agence France Presse is one of the largest news agencies in the world and is the largest in France. It is also one of the primary sources used by Western intelligence services to monitor the continent of Africa (Leetaru, 2010). LexisNexis describes the newswire in its Source Information directory as follows:

Agence France Presse is the world’s oldest news agency. Based in France, with staffers and stringers in 129 countries, AFP offers a unique perspective on the world's news. AFP's Europe coverage is outstanding, its reporting from Africa is renowned and its Latin American correspondence comprehensive. AFP also covers the Middle East, Asia and the Pacific Rim.

LexisNexis coverage of the newswire does not begin until May 1991. An extensive manual review of the source suggested it did not include an over-representation of coverage of domestic French affairs, focusing instead primarily on international coverage; earlier work by Phillip Huxtable came to a similar conclusion in comparing AFP and the English-language Reuters coverage of Anglo- and Franco-phone West Africa. This suggested there was no need to incorporate additional filtering to specifically remove articles discussing French affairs. In addition, Agence France Presse articles occasionally quote French governmental officials on their views towards an emerging situation, which would result in many relevant articles being discarded if keyword-based filtering was used to remove all articles mentioning France or French. While the newswire contains SECTION() metadata tags used to identify the major news desks such as sports and financial news, these are not always properly applied.

In addition, major sporting or financial news is often treated as general news, rather than being tagged under the appropriate section heading. An extensive manual review of a random selection of articles from each month was used to develop a lexicon of sports and financial-related keywords most commonly used in articles not properly tagged with the appropriate SECTION() tag. Thus, the following Boolean query was used to retrieve all articles from the Agence France Presse English file in LexisNexis:

```
NOT section(sports) AND NOT section(financial) AND NOT golf AND NOT baseball AND NOT football AND NOT basketball AND NOT tennis AND NOT cycling AND NOT cricket AND NOT rugby AND NOT volleyball AND NOT "formula one" AND NOT subject(sports) AND NOT subject(financial results) AND NOT subject(economic news) AND NOT subject(stock indexes) AND NOT industry(stock indexes)
```

Figure 2 plots the total number of articles per month available in the LexisNexis archive of the newswire, showing that the newswire underwent steady growth through a peak in March 2001 and steadily decreased its output over the subsequent decade through mid-2010. It has largely remained constant at an average of around 8,000 articles a month over the last three years. There are also several outages visible in the first two years of its appearance in LexisNexis, which is noted on its LexisNexis Source Information page. In all, LexisNexis records 2,135,896 articles totaling 661,009,337 words through September 2012, averaging around 309 words per article.

Figure 2: Articles per month Agence France Presse
Table 1 shows the top 25 countries most frequently discussed by Agence France Presse, ordered by the percent of all articles published by the service through September 2012 that mentioned each country. Here, any mention of the country or any city or other geographic landmark within the country was counted. Overall, there is a clear emphasis by Agence France Presse on Europe and the Middle East, with a particular focus on the United States and Great Britain.

<table>
<thead>
<tr>
<th>Country</th>
<th>All Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>10.44</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>4.79</td>
</tr>
<tr>
<td>France</td>
<td>4.31</td>
</tr>
<tr>
<td>Russia</td>
<td>3.56</td>
</tr>
<tr>
<td>China</td>
<td>3.41</td>
</tr>
<tr>
<td>Israel</td>
<td>3.10</td>
</tr>
<tr>
<td>Iraq</td>
<td>3.08</td>
</tr>
<tr>
<td>Germany</td>
<td>2.72</td>
</tr>
<tr>
<td>Japan</td>
<td>2.36</td>
</tr>
<tr>
<td>India</td>
<td>1.78</td>
</tr>
<tr>
<td>Iran</td>
<td>1.73</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>1.73</td>
</tr>
<tr>
<td>Pakistan</td>
<td>1.64</td>
</tr>
<tr>
<td>Italy</td>
<td>1.57</td>
</tr>
<tr>
<td>Egypt</td>
<td>1.33</td>
</tr>
<tr>
<td>Australia</td>
<td>1.31</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1.25</td>
</tr>
<tr>
<td>Spain</td>
<td>1.22</td>
</tr>
<tr>
<td>Turkey</td>
<td>1.21</td>
</tr>
</tbody>
</table>
2.0.2 Associated Press

The Associated Press is one of the largest news agencies in the world, operating 243 bureaus across the world. Unlike Agence France Presse, the Associated Press is operated as a cooperative, in which any story published by a member news agency is automatically redistributed and available for any other member to publish. LexisNexis describes the newswire in its Source Information directory as follows:

Founded in 1848, and now delivering news and photos in over 100 countries, The Associated Press sees itself as the oldest and largest news service in the world. The AP is a nonprot cooperative (i.e., a member-owned organization) with its roots in the newspaper industry. Regular members of the AP are obligated to report exclusively to the AP news that breaks locally, but might be of interest to the media elsewhere in the U.S. or overseas. This system gives the AP a news gathering reach well beyond what would be possible with only its staff resources. Coverage includes international news, national news (other than Washington-datelined stories), Washington news (only stories of national interest), business news, and sports.

The Associated Press newswire contains a wide assortment of news that heavily emphases domestic United States events, including local and regional newspaper coverage. Beginning in December 1978 the newswire added SECTION() metadata tags that allow the filtering of coverage to just national or international stories (prior to this date there was no choice but to download all coverage). The newswire also has a special designation of top news used to identify major breaking or important stories regardless of their geographic focus. Thus, the following Boolean query was used to retrieve Associated Press coverage from LexisNexis:

"top news" or section(international)

Figure 3 plots the total number of articles per month available in the LexisNexis archive of the newswire, reflecting the far lower volume of coverage compared with Agence France Presse. The sharp drop in coverage volume between November and December 1978 reflects the introduction of the new SECTION() metadata tag that allowed for retrieving just international articles. While the newswire underwent steady growth during the late 1990s, it has experienced a decade-long decline in its international coverage, stabilizing at around 1,600 articles per month over the last three years. In all, LexisNexis records 944,483 articles totaling 358,398,400 words through September 2012, averaging around 379 words per article. Table 2 shows the top 25 countries in terms of the relative percentage of all Associated Press coverage during this time that mentioned each. As with Agence France Presse, there is a strong emphasis towards European and Middle Eastern countries and a similar emphasis on French coverage.

<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>South Korea</td>
<td>1.15</td>
</tr>
<tr>
<td>Belgium</td>
<td>1.15</td>
</tr>
<tr>
<td>Syria</td>
<td>1.07</td>
</tr>
<tr>
<td>Canada</td>
<td>1.03</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>0.98</td>
</tr>
<tr>
<td>Lebanon</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Figure 3: Articles per month Associated Press

Table 2: Top 25 countries by percent of all Associated Press articles mentioning that country

<table>
<thead>
<tr>
<th>Country</th>
<th>All Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>13.35</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>5.32</td>
</tr>
<tr>
<td>Russia</td>
<td>4.51</td>
</tr>
<tr>
<td>France</td>
<td>3.39</td>
</tr>
<tr>
<td>Israel</td>
<td>3.36</td>
</tr>
<tr>
<td>Germany</td>
<td>2.74</td>
</tr>
<tr>
<td>Iraq</td>
<td>2.71</td>
</tr>
<tr>
<td>China</td>
<td>2.27</td>
</tr>
<tr>
<td>Japan</td>
<td>1.94</td>
</tr>
<tr>
<td>Iran</td>
<td>1.87</td>
</tr>
<tr>
<td>Italy</td>
<td>1.87</td>
</tr>
<tr>
<td>Egypt</td>
<td>1.50</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>1.44</td>
</tr>
<tr>
<td>Lebanon</td>
<td>1.44</td>
</tr>
<tr>
<td>Canada</td>
<td>1.27</td>
</tr>
</tbody>
</table>

http://semanticommunity.info/AOL_Government/Managing_Big_Data_For_Results
Updated: Wed, 23 Sep 2015 07:38:19 GMT
Powered by mindtouch™
Pakistan | 1.26  
India | 1.26  
Spain | 1.15  
Syria | 1.14  
West Bank | 1.12  
Saudi Arabia | 1.05  
Mexico | 1.04  
South Africa | 0.98  
Poland | 0.97  
Turkey | 0.97  

Those familiar with the Associated Press will likely question why the primary Associated Press newswire was used here, rather than the specialty Associated Press Worldstream newswire, which is exclusively focused on international news coverage. LexisNexis does in fact offer an archive of the Worldstream service that begins in October 1993 that is very comparable in terms of daily coverage volume to Agence France Presse. However, for unknown reasons the LexisNexis archive of Worldstream ends abruptly in July 2010, with coverage past this date exclusively carrying sporting results.

2.0.3 Xinhua

Xinhua is the official news agency of the People’s Republic of China and the largest news service in the country, operating 107 bureaus around the world. While it still retains its official role in promulgating the views and statements of the Communist Party, it has vastly expanded since its founding in the 1931 towards a general-purpose global news service competing with services like Reuters (Troianovski, 2010). LexisNexis describes the newswire in its Source Information directory as follows:

Xinhua is the authoritative source for information on Chinese government affairs, economic performance and Chinese views on world affairs. All Western news correspondents in Beijing rely on Xinhua’s English-language news report to keep abreast of Chinese affairs. The agency reports on Chinese affairs, including the economy, industry, trade, agriculture, sports and culture. Coverage includes diplomatic changes and extensive international reporting often from Africa or the Middle East. Xinhua also provides useful coverage of non-Chinese Asia.

As its description above suggests, Xinhua is extensively focused on domestic Chinese news, which would heavily overemphasize China over other countries. Through manual review of a random selection of articles from each month, it was determined that adding in exclusion keywords to drop those articles mentioning either China or Chinese removed domestic coverage with a minimal false positive rate. Unlike Agence France Presse, Xinhua coverage of international events uses quotes from Chinese officials far more sparingly, meaning this filtering criterion has a minimal impact on international coverage. Xinhua also has a dedicated financial newswire called Xinhua Economic News Service, separating Xinhua’s extensive coverage of the financial markets from the Xinhua General News Service newswire used here. It does not, however, offer the SECTION() tags used with

http://semanticommunity.info/AOL_Government/Managing_Big_Data_For_Results
Updated: Wed, 23 Sep 2015 07:38:19 GMT
Powered by mindtouch™
Agence France Presse and Associated Press coverage to filter out sports-related coverage, necessitating the use of additional keyword filters. Thus, the following Boolean query was used to retrieve Xinhua coverage from LexisNexis:

```
NOT china AND NOT Chinese AND NOT olympic AND NOT snooker AND NOT boxing AND NOT hockey AND NOT marathon AND NOT motorcycling AND NOT soccer AND NOT handball AND NOT cycling AND NOT tennis AND NOT world cup AND NOT basketball AND NOT wrestling match AND NOT wrestling score AND NOT iceskating
```

Figure 4 plots the total number of articles per month available in the LexisNexis archive of the newswire. The near-tripling of coverage between December 1998 and November 1999 reflects the US involvement in Iraq during this period, which attracted a singularly massive volume of coverage from Xinhua. The service also has two major outage periods in LexisNexis, from April 1995 to June 1996 (inclusive) and April 2008 to October 2008 (inclusive), so those are removed from consideration for all analyses. In all, LexisNexis records 1,699,442 articles totaling 332,043,292 words through September 2012, averaging around 195 words per article. Table 3 shows the top 25 countries in terms of the relative percentage of all Associated Press coverage during this time that mentioned each. As with Agence France Presse and the Associated Press, there is a strong emphasis towards European and Middle Eastern countries.

**Figure 4: Articles per month Xinhua**

![Figure 4: Articles per month Xinhua](image)

**Table 3: Top 25 countries by percent of all Xinhua articles mentioning that country**

<table>
<thead>
<tr>
<th>Country</th>
<th>All Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>10.76</td>
</tr>
<tr>
<td>Israel</td>
<td>3.81</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>3.34</td>
</tr>
<tr>
<td>Russia</td>
<td>3.27</td>
</tr>
<tr>
<td>Iraq</td>
<td>2.96</td>
</tr>
<tr>
<td>France</td>
<td>2.43</td>
</tr>
<tr>
<td>Country</td>
<td>Value</td>
</tr>
<tr>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>Japan</td>
<td>2.32</td>
</tr>
<tr>
<td>Pakistan</td>
<td>2.01</td>
</tr>
<tr>
<td>India</td>
<td>1.98</td>
</tr>
<tr>
<td>Egypt</td>
<td>1.94</td>
</tr>
<tr>
<td>Iran</td>
<td>1.94</td>
</tr>
<tr>
<td>Germany</td>
<td>1.92</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>1.88</td>
</tr>
<tr>
<td>Thailand</td>
<td>1.53</td>
</tr>
<tr>
<td>Philippines</td>
<td>1.44</td>
</tr>
<tr>
<td>South Africa</td>
<td>1.42</td>
</tr>
<tr>
<td>Australia</td>
<td>1.41</td>
</tr>
<tr>
<td>Turkey</td>
<td>1.37</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1.34</td>
</tr>
<tr>
<td>Syria</td>
<td>1.26</td>
</tr>
<tr>
<td>Lebanon</td>
<td>1.14</td>
</tr>
<tr>
<td>Nigeria</td>
<td>1.08</td>
</tr>
<tr>
<td>West Bank</td>
<td>1.07</td>
</tr>
<tr>
<td>Kenya</td>
<td>1.07</td>
</tr>
<tr>
<td>Italy</td>
<td>1.06</td>
</tr>
</tbody>
</table>

### 2.0.4 Comparing the Sources

This section will briefly compare some of the characteristics of the three news sources in order to tease apart their mutual differences and explore whether predictive features found in one source are universal across the others; additional details can be found in Leetaru [2013]. In total across the three sources, 4,779,821 articles were processed in the course of this dissertation, totaling 1,351,451,029 words.
Figure 5 shows the Z-scored (standard deviations from mean) plots for all three sources overlapping their relative growth and decay patterns. Figure 6 shows the combined monthly article volume across the three sources demonstrating in particular the significant mutual growth during the 1990s. The three sources, however, are poorly correlated in their temporal profiles. Even restricting the analysis to only overlapping periods of time, the monthly coverage volume of Agence France Presse has a Pearson correlation of \( r = 0.27 \) with Xinhua and \( r = 0.35 \) with Associated Press, while the Xinhua and the Associated Press are correlated at \( r = 0.03 \). While weak, Agence France Presse is correlated with the other two sources at \( p < 0.0005 \), indicating high statistical significance, with Agence France Presse and the Associated Press being the only two sources not to have a statistically meaningful correlation. In terms of geographic emphasis, the three sources are tightly aligned, with Agence France Presse being correlated at \( r = 0.97 \) in terms of the relative percentage of eachs coverage dedicated to each country, while Agence France Presse is correlated at \( r = 0.94 \) with Xinhua. Xinhua and the Associated Press are correlated at \( r = 0.95 \). All three are therefore at the highest level of statistical significance \( (p < 0.0005) \), indicating there are no significant differences in their respective geographic focus.

**Figure 5: Z-scored articles per month comparing Agence France Presse, Associated Press, and Xinhua**

**Figure 6: Combined articles per month across Agence France Presse, Associated Press, and Xinhua**

### 2.0.5 Additional Post Filtering

Despite careful construction of the queries used to retrieve each newswire from LexisNexis, including extensive manual review of random selections of content to develop exclusion keywords, a non-insignificant volume of sports and financial coverage was still retained. Such coverage can create complications for the event coding process in that it often contains language very similar to that used to describe violent political events, such as two sports teams battling it out or a companys stock price "under siege" [Schrodt and Van Brackle, 2013]. Thus, after all matching articles were downloaded from each newswire, a second extensive manual review was performed across the combined content pool to filter out any remaining sports or economic-related coverage. While this filter may eliminate certain economic-related articles which might reflect or drive public opinion (such as an economic boom or bust), incorporating this filter dramatically reduced the false positive rate of Tabari. Thus, the following Boolean query was applied in a post-processing stage after the content was downloaded, but before it was made available for secondary processing.
NOT boxing AND NOT hockey AND NOT marathon AND NOT motorcycling AND NOT soccer AND NOT handball AND NOT cycling AND NOT tennis AND NOT worldcup AND NOT world cup AND NOT basketball AND NOT wrestling match AND NOT wrestling score AND NOT icestaking AND NOT ice staking AND NOT skiing AND NOT football AND NOT coach AND NOT hockey AND NOT box oce AND NOT snooker AND NOT cricket AND NOT game console AND NOT gaming console AND NOT tv show AND NOT bond market AND NOT currency trade AND NOT closed up AND NOT closed down AND NOT industrial average AND NOT nasdaq AND NOT dow jones AND NOT halftime AND NOT half time AND NOT the game AND NOT stocks declined AND NOT market declined AND NOT inflation AND NOT interest rate AND NOT regional growth AND NOT car sale AND NOT truck sale AND NOT midsize car AND NOT inflation AND NOT singer AND NOT teammate AND NOT team mate AND NOT freethrow AND NOT free throw AND NOT show times AND NOT athletic AND NOT free throw AND NOT touchdown AND NOT the season AND NOT rebounds AND NOT quarterback AND NOT point guard AND NOT fourth quarter AND NOT on the road AND NOT season high AND NOT diet AND NOT title bid AND NOT mixed doubles AND NOT bowl game AND NOT retail price AND NOT book review AND NOT garden AND NOT goalkeep AND NOT goal keep AND NOT mega million AND NOT megamillion AND NOT mega-million AND NOT lottery game AND NOT lottery winner AND NOT ticket sale AND NOT lottery jackpot AND NOT baseball AND NOT golf AND NOT growth outlook AND NOT the dollar AND NOT bank index AND NOT nf l AND NOT nh l AND NOT nba AND NOT sports AND NOT championship AND NOT entertainment

This leads to the following processing pipeline used to construct the event database used here:

1. All relevant content from each newswire is downloaded from LexisNexis Academic Universe using the source-specific query.

2. Each article is subjected to the additional post-processing Boolean query to drop remaining sports- and financial-related news coverage.

3. Each article is subjected to fulltext geocoding from Leetaru [2012] to identify and disambiguate all geographic references contained in each article.

4. The Tabari system is applied to each article in full-story mode to extract all events contained anywhere in the article and the Tabari geocoding post-processing system is enabled to georeferenced each event back to the specific city or geographic landmark it is associated with.

5. The final list of events for each newswire is internally deduplicated. Multiple references to the same event across one or more articles from the same newswire are collapsed into a single event record. To allow the study of each newswire individually, events are not deduplicated across newswires (externally deduplicated).

In all, there were 28,877,172 events identified and coded by Tabari from the three news wires: 14,433,748 from Agence France Presse, 7,811,104 from the Associated Press, and 6,632,320 from Xinhua. On average, there were 7 events per article in Agence France Press, 8 from Associated Press, and 4 from Xinhua. However, since the three sources have very different average article lengths, when calculating the average words per event, the results are far more even: 46 words per event on average for both Agence France Presse and the Associated Press, and 50 words per event for Xinhua.

Figures 7, 8 and 9 show the total number of events per month for each news source in the four Quad Classes of Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material Conflict. The gap in Associated Press events for 1992 is due to a technical error with the content downloaded for that year that prevented it from being properly processed, this has been corrected in the final data set. Both Agence France Presse and Xinhua have very similar volumes of Material Cooperation and Verbal and Material Conflict events, making it hard to distinguish the relative change over time in those three categories, while their Verbal Cooperation...
events have clear temporal signatures. Associated Press coverage, seen in Figure 8, however, has strong stratification among the four classes, making it clear that they are closely aligned. Indeed, the four series are correlated at between $r = 0.80$ and $r = 0.97 (p < 0.0005)$ for all three news wires. Finally, Table 4 shows the relative breakdown of all events into the four Quad Classes for each newswire. It is clear that Verbal Cooperation events are the most common, followed by Material Conflict, Verbal Conflict, and Material Cooperation. In all, across the three sources, 60.56% of events were Verbal Conflict, 17.34% were Material Conflict, 13.12% were Verbal Conflict, and 8.99% were Material Cooperation.

Figure 7: Agence France Presse events per month by Quad Class

Figure 8: Associated Press events per month by Quad Class

Figure 9: Xinhua events per month by Quad Class
Table 4: Breakdown by percent of all events in each newswire in each Quad Class

<table>
<thead>
<tr>
<th>Quad Class</th>
<th>AFP</th>
<th>AP</th>
<th>Xinhua</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Cooperation</td>
<td>60.35</td>
<td>54.63</td>
<td>68.02</td>
</tr>
<tr>
<td>Material Conflict</td>
<td>17.37</td>
<td>20.75</td>
<td>13.22</td>
</tr>
<tr>
<td>Verbal Conflict</td>
<td>13.75</td>
<td>14.45</td>
<td>10.16</td>
</tr>
<tr>
<td>Material Cooperation</td>
<td>8.53</td>
<td>10.17</td>
<td>8.60</td>
</tr>
</tbody>
</table>

3 Automated Coding Engines

In the field of event data, multiple independent tests [King and Lowe, 2004, Schrodt and Gerner, 1994] have shown that machine coding is comparable in accuracy to human coding. Furthermore, the human coding accuracy in some of those tests is quite low: King and Lowe [2004] find the accuracy on the individual VRA event codes alone (Table 2, pg 631) -- not the complete record with source and target identification -- is in the range 25% to 50% for the detailed codes and 55% - 70% for the major categories. Similarly, Mikhaylov et al. [2012] show that the reliability of the human coding in the widely-used Comparative Manifestos Project is less than half what is commonly reported, and for some indicators drops as low as 25%; Ruggeri et al. [2011] show similar problems in the coding of governance events in UN peacekeeping.

An extensive recent body of psychological work -- see Baumeister and Tierney [2011] for a popular treatment -- indicates that the sustained decision-making required for human coding presents an almost perfect storm for inducing fatigue, inattention, and a tendency to use heuristic shortcuts. These psychological costs are far more deeply rooted than previously assumed, and can only be reduced, not eliminated, by improved coding protocols, training, coder selection and supervision. The human brain was simply never intended for the tasks we impose on coders.

Improvements in computer software and hardware, meanwhile, have made the process of analyzing large bodies of text much more efficient, and the field has matured with the development of a common set of methodologies with well-tested characteristics. Automated coding is completely transparent, without the unreproducible subjective elements of human coding. Moreover, once the source texts have been prepared, recoding to account for new theoretical or technological components can be done quickly and efficiently.

Automated coding also opens the possibility of near-real-time (NRT) coding, which can scale to arbitrarily large sets of source texts. In the early phases of the ICEWS project, an implementation of Tabari on a small cluster computer reliably coded 26-million news reports in six minutes, resulting in about 3-million events. In contrast, sustained human coding projects, once one accounts for training, retraining, replacement, cross-coding, recoding due to effects of coding drift and/or slacker-coders and so forth, usually code about six events per coder-hour and -- like the labor requirements of a string quartet -- have changed little over time. The arithmetic is obvious: 3-million events from six minutes of automated coding, or 500,000 labor-hours of manual coding, probably costing on the order of $10-million when labor and administrative costs are taken into effect.

3.1 Event Coding
3.1.1 TABARI

The core engine for the event coding was the open-source Tabari program. While this program has been in use for a number of years, the GDELT project provided an incentive for several enhancements

- We improved the ability of Tabari to automatically assemble codes from combinations of a named actor and an generic agent. For example "Philippine soldiers" will automatically generate the code PHLMLIL, whereas "The Philippine Secretary of Agriculture" will automatically generate the code PHLGOV. Earlier dictionaries had done this directly, with separate dictionary entries for, say, "Australian police," "Cambodian police," "Chinese police" and so forth. The new system is both faster in terms of the dictionary size and much more efficient. This allows the coding of both generic agents such as "police," "soldiers," "demonstrators" and the like, as well as named individuals where we have the title in the dictionary but not the individual person. In support of this new facility, we also increased the size of the :agents dictionary considerably based on WordNet.

- The Tabari CONVERT AGENTS facility generalizes this further and allows agents which do not have an associated actor to be converted to actors and their agent codes (rather than actor+agent codes) are used. This is allows a sentence such as

Students and police fought in the Egyptian capital

being coded as

EDU fought COP

or the sentence

Sudanese students and police fought in the Egyptian capital

being coded as

SUDEDU fought COP

This allows coding where the location of the event can be determined by processing with other programs. The use of this option increased the number of coded events by about 22%.

- A new facility was added to Tabari that made the source and target actors optional: this allows both for events such as general statements where there is not an explicit target, and also the coding of situations where an event is implied by the structure of the sentence, but the system cannot find the appropriate source or target in the dictionaries. Sentences coded in this manner contain a default code so that they can be easily eliminated at a later processing stage if a researcher does not want to include this information, but are useful in aggregations where the identity of the [typically] target is not important.

3.1.2 Dictionaries

The Tabari system has very extensive open-source dictionaries for the identification of political actors. Central to these is the 32,000-line CountryInfo.txt (http://eventdata.psu.edu/software.dictionaries.html), a general purpose lex intended to facilitate natural language processing of news reports and political texts. This covers about 240 countries and administrative units (e.g. American Samoa, Christmas Island, Hong Kong, Greenland); fields include adjectival forms and synonyms of the country name, the capital city and cities with populations over 1-million, regions and geographical features (WordNet meronyms), leaders from http://rulers.org and members of government from the CIA World Leaders open database.

CountryInfo.txt was supplemented by a number of lists of actors such as MNCs, NGOs and IGOs, and militarized non-state actors such as al-Qaeda and the Lords Resistance Army. Under NSF funding, the CAMEO actor coding system has been enhanced...
with a very extensive religion typology, the CAMEO Religious Coding Scheme (CAMEORCS), which has a hierarchical coding of about 1,500 religious denominations, and an ethnic group coding scheme with about 650 ethnic groups. The formal names of these have been implemented as Tabari dictionaries, though we have not had the opportunity to evaluate how frequently this translates into actual actor codes.

As a consequence of these enhancement, the actors dictionaries now have around 60,000 entries, compared to the 1,000 or so entries typical in earlier KEDS work; this is further supplemented with about 1,500 agents (common nouns), which can also be treated as actors either in combination or alone.

The verb coding dictionary was the standard one we have been using since about 2010—this was a point when regular noun and verb forms were added to Tabari—which proved remarkably robust in a number tests we did in the early phases of ICEWS. As noted in the conclusion, we will soon be shifting to a new dictionary which employs a extensive set of WordNet synonyms both in the verbs themselves and also in common phrases where synonym sets might be encountered, such as the names of monetary currencies and the names of weapons. We have identified about a dozen such categories and have largely incorporated these into the dictionaries, but this was not used in this version of GDELT.

3.2 Location and Tone

The raw Tabari output files are geolocated by a post-processing system. The first system takes the verb and actor word sentence offsets (that is, the identity of the word in the sentence that was the start of each actor and verb mention) and converts them back into their locations in the original article. This is then cross-walked against the list of all geographic locations found in the text, and the closest location to each of the actor and verb mentions is assigned to it. While this is quite primitive, it works surprisingly well and was the approach used to map Wikipedia. The "tone" of each article is measured using the tonal algorithm from Shook et al. [2012].

3.3 "The Pipeline"

3.3.1 Pre-processing

- First the article is cleaned up, any textual URLs, phone numbers, email addresses, non-ASCII characters, and other material is removed.
- The article is then run through full-text geocoding that automatically identifies and disambiguates all geographic references in the text, resolving them to their centroid lat/long coordinates, using Leetaru (2012).
- The article then goes through a preprocessing pipeline that creates four versions of the text:

  - RAW: The text is kept as-is with no change.

  - CITY TO COUNTRY NOPERSON: All mentions of cities and other geographic landmarks are replaced with the name of the country the city/landmark is located in. Thus, "Soldiers marched in Cairo today" will be replaced with "Soldiers marched in Egypt today." While the Tabari :actors dictionary contains major city names, it contains only the largest cities, and a mention of soldiers attacking a small village will cause Tabari not to find a country affiliation for the victim actors, while this interpolation will replace that small village with its corresponding country name, causing Tabari to find a match. Person names remain as-is.

  - PERSON TO COUNTRY: For each person's name found in the text, the closest location to the rst mention of the person's name is then assigned to that person and all mentions of the person's name are then replaced with that location's country name. Thus, a mention of "Egyptian Minister of Foreign Aairs Mohamed Orabi attended the summit yesterday. While
he was there, Orabi pledged support for..." This is essentially a special geographically-centered form of entity dereferencing. This version dramatically improves matching of diplomatic events in which a political leader is referred to throughout an article, but the leader is not significant enough to have his or her own entry in the Tabari :actors list. In addition, all mentions of cities and other geographic landmarks are replaced with "CityName, CountryName."

- FULL: In this version, each person name is associated with its corresponding location as in PERSONTOCOUNTRY, but all mentions of the person's name are replaced with the form of "PersonName of CountryName" (adding the country name and inserting the word "of" between the person's name and the country name). All mentions of cities or landmarks are converted to the format "City- Name, CountryName."

3.3.2 Coding

Each of the text versions above causes Tabari to catch different events and significantly improves Tabari's ability to capture complex events, since it essentially "rewrites" each sentence several different ways for the Tabari grammar to catch it: these rules make the sentence "Tabari-friendly."

• For each of the versions above, the text is processing through Tabari twice, once with the standard settings and once with COMMA: OFF set in the :options file. Through extensive testing, it appears that on more complex articles, this dual-pass coding process yields a substantial increase in the number of recovered events without yielding a measurable number of false positives.

• Tabari is designed to process only a single sentence at a time, so the above text is then converted to the per-sentence input format needed by Tabari. A special sentence-splitting algorithm is used that is robust to periods occurring in different contexts such as abbreviations and typographical errors, and multiple blank lines being used as paragraph separators, etc. In addition, an optimization process uses a fast surface scan of the sentence to determine whether it contains matches of at least one entry from the Tabari :verbs list. It does not attempt to determine if there would be an actual match based on a grammatical parse, but simply checks whether it would even be possible for the sentence to have a match. This is extremely fast and eliminates 70-80% or greater of all sentences, vastly speeding up the rest of the processing pipeline (both the input/output requirements of managing all of those sentences and Tabari's processing time).

• While Tabari ordinarily is used only to code the first sentence of each article, here it is used to code all sentences to ensure it codes all of the surrounding contextual events. This vastly improves its recognition of complex situations as well in that while the lead sentence may be too grammatically complex in some cases for Tabari to process, those events will be repeated later in the text, so it ultimately codes all of the events. Events coded in the first lead paragraph of the article are marked using the RootEvent flag so that users can select only these events as desired. In addition, TIME SHIFTING is enabled to ensure that mentions of past events are coded with the proper date.

• Historically Tabari would code a blank actor if there was AGENT information available, but no actor (such as gunmen or police). The new CONVERT AGENT = TRUE functionality of Tabari is used here so that in the absence of an available actor, Tabari will code any available AGENT mentions as the actor instead. This yields an average of 22% additional events that were previously lost.

3.4 One-A-Day Filtering

Following the protocols used in most of the Tabari-based research, the major post-processing step is the application of a "one-a-day" filter, which eliminates any records that have exactly the same combination of date, source, target and event codes. This is designed to eliminate duplicate reports of events that were not caught by earlier duplicate news report filters. In our work on the
Levant data set, this fairly consistently removes about 20% of the events; the effect on the ICEWS data may be somewhat higher due to the use of a greater number of sources.

In areas of intense conflict—where multiple attacks could occur within a single dyad in a single day—this could eliminate some actual events. However, these instances are rare, and periods of intense conflict are usually obvious from the occurrence of frequent attacks across a month (our typical period of aggregation), and do not require precise measures within a single day. Periods of intense conflict are also likely to be apparent through a variety of measures—for example comments, meetings with allies, offers of aid or mediation—and not exclusively through the attacks themselves.

Each record is then converted into a unique identifier key that concatenates the actor, action, and location fields. This is then checked against the list of all existing events in the database: if the event already exists, the previous event record has its NumMentions and NumSources fields updated accordingly, otherwise a new record is inserted into the database. Traditional Tabari event deduplication is performed at the country-day level, however in high-conflict regions like Syria, it is necessary to incorporate the city-level information into the event record to preserve the locations of each riot and to separate riots in different parts of the country into distinct event records.

### 4 Visualizations

#### 4.1 Quad Count Time Series for Selected Dyads

The figures in Table 5 shows quad-count time series plots for four active dyads: Israel-Palestine, Israel-Lebanon, China-Taiwan and India-Pakistan. With the exception of the Israel-Lebanon case, the most apparent feature of these is the increase in the number of events over time; this scaling issue also masks signicant variation in each series in the first half of the data.

**Table 5: Quad Count Time Series for Selected Dyads**
4.2 District-level Conflict in Afghanistan

Yonamine [2013, chapt. 4] uses GDELT to develop models which predict district-level conflict in Afghanistan, finding that the more finely-aggregated district data produces more accurate predictions than province-level and national-level data. Figure 10 shows the distribution of the data over time: note that this looks similar to maps that have been produced using the Wikileaks data—the extent to which this can be used in research still very much in legal limbo—but was produced entirely from open-source, contemporaneous sources.

Figure 10: The Number of Material Conflict events per Afghani District from 2001 to 2012 (Yonamine 2013)

4.3 Comparison of GDELT and NGO-based data for City-level violence in Syria

Yonamine [2013, chapt. 1] compares city-level reports of violent events in GDELT and the level of violence reported by a Syrian-based NGO: these two sources are shown in Figures 11 and 12. Despite the fact that Syria is a particularly difficult conflict to cover—on a typical day, the number of deaths reported in the major international media are less than half those reported in an NGO source such as http://syriashuhada.com—the shape of the curves for Aleppo and Homs are almost identical between the two distinct sources.
Figure 11: The Number of GDELT-derived Violent Events in Homs and Aleppo from January 2012 through June 2012 (Yonamine 2013)

Figure 12: The Number of NGO-reported Violent Events in Homs and Aleppo from January 2012 through June 2012 (Yonamine 2013)

4.4 GDELT: 29 January 2011

Figures 13 through 15 show three different visualizations of the GDELT world on 29 January 2011, in the middle of the Egyptian Spring.

Figure 13 uses modification of a special heatmapping process developed by the cybergis group ([http://www.psc.edu/media/training/XH...ulturomics.pdf](http://www.psc.edu/media/training/XH...ulturomics.pdf); Shook et al. [2012]). For each Action location it evaluates the total number of events at the point and the average Goldstein score [Goldstein, 1992] of all of those events. As such, it gives a rough heatmap showing the major areas of positive and negative events, emphasizing the underlying clusters. This technique shows the major groupings of positive and negative actions, filtering them from the background landscape of events.
Figure 13: GDELT for 29 January 2011 based on number of events and Goldstein scores

Figure 14 shows the same day, but uses the raw event data. It sizes the dot by the NumberMentions field; the colors are based on the Goldstein scale. For clarity, NumberMentions is restricted to a maximum, so all major events are around the same size.

Figure 14: GDELT map for 29 January 2011 based on number of mentions of event and Goldstein scores

Figure 15 puts dots for the locations of both Actor1 and Actor2 and then draws a line between them colored by the average Goldstein score of the event. This shows which areas of the world are being connected to what other areas of the world.

Figure 15: GDELT network map for 29 January 2011 based on dyadic links and average Goldstein scores

5 Comparison to Other Data Sets
5.1 KEDS/Reuters Levant

The KEDS Levant data set ([http://eventdata.psu.edu/data.dir/levant.html](http://eventdata.psu.edu/data.dir/levant.html)), which covers April-1979 to December-2011, is the only available event data set that is comparable to GDELT in terms of temporal coverage. While it is also coded with Tabari, the dictionaries for this set were extensively configured over a period of years under a number of NSF-funded projects from about 1990 to 2005--the Tabari :verbs dictionaries used for the global coding were largely developed on Levant texts--whereas GDELT uses generic global :actors and :agents dictionaries. In this respect, comparing the two datasets is an assessment of whether the GDELT approach using global dictionaries produces results comparable to the laboriously hand-tuned KEDS data.

The larger difference between the two sets, however, is in the sources. The Levant data consists of two single-source datasets, one based in Reuters, the other--covering only 2000 to the present--on AFP. GDELT, of course, uses a much larger set of sources--though it does not directly incorporate Reuters--so any differences between the two sets will be a function of both the dictionaries and the sources. To maximize coverage, we will look at the Reuters series, and focus on the two dyads in the series with the most activity: Israel->Lebanon (LBN) and Israel->Palestine (PSE).

Figures 16 and 17 show scatterplots for the entire period for two of the quad counts: verbal cooperation and material conflict. Note that the horizontal scale for GDELT is 10-times that of the vertical scale for Reuters: unsurprisingly, the multiple news sources used by GDELT produce, on average, a far higher density of data. The correlations between the two sets are reasonably high, though the scatters are quite wide.

**Figure 16: Comparison of KEDS/Reuters and GDELT Quadcounts for Israel -> Lebanon**

As we saw in Figure 1, the density of data in GDELT varies substantially over time, and shows an exponential increase after 2002 or thereabouts. Furthermore, the activity within these dyads has also changed, particularly with the two
intifadas and various abortive peace processes in the Israel->Palestine dyad. Consequently, the relationship between the two data set could plausibly vary with time, so we analyzed the three intervals shown in Table 6 separately.

Table 6: Hypothesized intervals in Levant data

<table>
<thead>
<tr>
<th>Interval</th>
<th>Years</th>
<th>ISR-&gt;LBN</th>
<th>ISR-&gt;PSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1979-1987</td>
<td>Israeli invasion</td>
<td>Camp David</td>
</tr>
<tr>
<td>B</td>
<td>1988-1999</td>
<td>Israel-Hizbollah conflict</td>
<td>First intifada, Oslo</td>
</tr>
<tr>
<td>C</td>
<td>2000-2011</td>
<td>Cold peace</td>
<td>Second intifada, PA/ Hamas split</td>
</tr>
</tbody>
</table>

These results are shown in the gures in Table 7—once again note that the GDELT and Reuters counts are on different scales—which shows a very clear pattern of low correlations—nearly zero in the case of Israel -> Palestine verbal cooperation—in the post-2000 period, but relatively high correlations pre-2000. In several cases, in fact, the pre-2000 correlations probably exceed those that would be expected from human inter-coder reliability, but in all cases are reasonably high. As emphasized in Figure 18, the low correlations are strongly influenced by a large number of points where GDELT produces events but the Reuters set does not. As far as we can tell, the differences between the sets are mostly due to the differences in the presence of events, rather than differences in what is being reported, though we have not done sufficient tests to rule this out entirely.

Table 7: Comparison of KEDS/Reuters and GDELT Quadcounts by Time Period
This comparison—along with Figure 1—illustrates what is clearly going to be a major challenge in using GDELT for long time-series studies: the post-2000 increase in the frequency of reports. As shown in Table 5, this does not affect all dyads equally: except for a spike in 2006 that corresponds to the 34-day war involving Israel, Lebanon, and Hezbollah in the ISR/LBN data (upper right figure in the table), the period after 2000 for this dyad doesn’t look all the different than the period prior to 2000, and the drop-off in the correlations in Table 7 are not nearly as high for Lebanon as for Palestine. Nonetheless, caution is very much advised on this issue, and given that the international news reporting environment is still changing, it is likely to be some time before this situation stabilizes.

5.2 ICEWS

A second comparison will be with one of the later ICEWS data sets. This was labeled "Release 28" and was coded either with Jabari or possibly some version of Jabari-NLP; this would have been one of the last versions of the Asian ICEWS data before the project switched to development of the global W-ICEWS set. The data nominally go to Mar-2011 but the last couple of months have very low counts and may have been incomplete, so the series, which begins in Jan-1998, was truncated at Dec-2010. ICEWS is based on a large number of sources from Factiva.

While the ICEWS forecasts primarily focus on internal conflict, the coding dictionaries work for international interactions as well—in fact—international crisis" is one of the ICEWS "events of interest" indicators—and so we will compare the two data sets on four international dyads: China->Taiwan, India->Pakistan, South Korea->North Korea and USA->Japan.

Figure 19 shows the China->Taiwan comparison. The correlations—[0.024, -0.16, 0.620, 0.275] in the order [VERCP, MATCP, VERCF, MATCF]—are relatively low compared to the KEDS/GDELT comparison, except for the VERCF counts. As with the KEDS/GDELT comparison, GDELT has a higher density of events, though it tends to be only two to three times higher. VERCF is again the exception to this, and the two datasets have roughly equal densities here.
These are not particularly good correlations. Three factors may be contributing to the divergence. First, most of the GDELT sequence being compared here is in the post-2000 period when GDELT is experiencing an exponential increase in density; as we saw in Section 5.1, this period had much lower correlations in the KEDS/GDELT comparison. Second, GDELT includes Xinhua, which ICEWS does not include, and Xinhua may be throwing off the totals when compared to the international sources. This in particular might explain why the VERCF (verbal conflict) indicator has the highest correlation: that would be consistent with Xinhua being used as a tool of the Chinese government's generally belligerent foreign policy towards Taiwan, and those policy pronouncements, in turn, would be monitored by the international media.

Finally, there appears to be serious discontinuity in the latter part of the ICEWS sequence, which drops from reporting tens of events per month to ones of events. Using V ERCP >=30--that is, an average of at least one event per day--as a threshold, the correlations improve considerably, as shown in Figure 20. The vector of correlations here is [0.525, 0.350, 0.804, 0.583], which is more or less in line with the KEDS/GDELT comparison for the pre-2000 period.

Figures 21, 22 and 23 show scattergrams for the remaining dyads: these generally show patterns similar to those seen in the China->Taiwan case. Figure 21 for India->Pakistan shows the same pattern of a high correlation [0.623] for VERCP--though this is clearly
inflated by an outlying point which is similar in both data sets and relatively low correlations [0.16 to 0.25] for the remaining counts; the ratio of the GDELT to ICEWS counts is again in the range of two to three. South Korea->North Korea, Figure 22, has higher correlations, in the range [0.55-0.75] except for MATCP, though again these are inflated by outliers. The ratio of GDELT to ICEWS counts is substantially higher here, in the range of five to ten; again it is possible that Xinhua accounts for the difference. Finally, the USA->Japan dyad, Figure 23, has very low correlations, once again strongly influenced by the very low counts on all of the ICEWS indicators except VERCP.

**Figure 21: Comparison of ICEWS and GDELT Quadcounts: India->Pakistan 1998-2010**

![Figure 21: Comparison of ICEWS and GDELT Quadcounts: India->Pakistan 1998-2010](http://semanticommunity.info/AOL_Government/Managing_Big_Data_For_Results)

**Figure 22: Comparison of ICEWS and GDELT Quadcounts: South Korea->North Korea 1998-2010**

![Figure 22: Comparison of ICEWS and GDELT Quadcounts: South Korea->North Korea 1998-2010](http://semanticommunity.info/AOL_Government/Managing_Big_Data_For_Results)
These comparisons clearly need to be explored in further detail, the most critical issue being further analysis to ascertain whether the difference in the counts is due to ICEWS being more selective--at various points, the project was working on making the coding very sensitive to false positives, particularly on conflict events--or whether GDELT is capturing more detail, particularly in the post-2000 period, because it is using a wider variety of web-based sources. A systemic drop in reports in ICEWS at the same time GDELT is experiencing an exponential increase in reports would, of course, be a perfect storm for the sequences not correlating.

To further explore this possibility, we ran two-group t-tests on the 32 sequences (quad categories x dyad x GDELT/ICEWS), splitting the series at Jan-2005. Twelve of the sixteen tests showed significant differences ($p < 0.05$) for ICEWS, and the same ratio occurred for GDELT. However, the direction of these were quite different between the two data sets: of the twelve significant differences in ICEWS, ten were positive (the counts in the pre-2005 were higher), whereas for GDELT only four were positive. Finally, except in the South Korea->North Korea (where all directions were the same) in all of the cases where both t-tests were significant, the changes were in opposite directions. All of this would suggest that much of the difference between the sets is due to changes in the baseline frequencies rather than the coding of specific events.

All factors being equal (though in this case, they clearly aren’t equal due to differences in the source texts), the ICEWS data should be more accurate than the GDELT data because Jabari-NLP has some clear advantages over Tabari [Schrodt and Van Brackle, 2013].

Our sense, however, is that there is more going on here than just the difference in coding engines, since those accuracy improvements are unlikely to have exclusively resulted in the elimination of events. At this point, it would be very helpful if we had additional updated and documented sets of ICEWS and W-ICEWS data, but at present these do not seem to be available. Until we further explore the characteristics of the ICEWS data, this comparison should not be considered definitive.

6 Future Plans and Collaboration Opportunities

The authors are already well underway on work for the next release of this event dataset, which include an expanded event taxonomy, an array of new event attributes (including estimates of the number killed and injured), and a completely new event processing system based on a flexible new natural language processing system. In addition, the next release will extend coverage back to 1800 for all countries.

As noted in Section 3.1.2, we have new versions of the verb dictionaries that are organized around WordNet synonym sets for both the core verbs and common noun sets such as those using units of currency. These should make the dictionaries more...
robust, since a general phrase such as "Nation X has agreed to provide Nation Y currency-million in development aid" will need to be entered only once, rather than separately for each currency. All the work has been done on these; we just need to do some additional formatting.

While Tabari now has the ability to handle such "synsets," Schrodt along with Michael Ward and Jay Ulfelder recently received NSF-funding which will provide for the development of a completely new coder written in Python, at which point we will retire Tabari. This programming, which will be developed on the open-collaboration platform GitHub, was motivated by several issues:

- While thoroughly debugged--Tabari coded the 200-million events of GDELT without crashing--the code base is about twelve years old and written in the computer language C++, which no longer has a large (or young) programming community; Python is much more suited than C++ for text processing--we are hoping the choice of Python will reduce the remaining required programming to a point where this will be a summer-length project--and in a number of experiments we have done with very large scale textual databases, seems sufficiently fast (and in any case, we can always run large jobs on a cluster computer)
- Based on the experience of the Lockheed Jabari-NLP system developed for the latter phases of ICEWS [Schrodt and Van Brackle, 2013], we will be incorporating a number of open-source natural language preprocessors so that the system can work on a parsed representation of the text, rather than continuing the "shallow parsing" approach of Tabari and Keds. This should be particularly helpful in improving the accuracy of target actor identification.
- The various GDELT pre-processing steps described in Section 3.3.1 will be also incorporated into a Python-based pre-processing suite which can be used for standard processing.

We are also hoping to leverage efforts elsewhere in the social science data community for the further standardization of non-state actor codes: while we have developed a set of religious and ethnic codes in CAMEO, one of the primary advantages of automated coding is that recoding using a new (or multiple) systems is quite easy once dictionaries have been developed, whereas this was nearly impossible, and we are certainly open to other suggestions. In addition, we will be extending the CAMEO coding system--and the dictionaries--to incorporate new categories of events relating to finance, criminal activity, disease, and natural disasters. Finally, under the new NSF funding we are expecting to set up one or more "crowd-sourcing" platforms--most likely within expert communities (and their conscripted students) rather than for web denizens in general--that will allow collaborative development of dictionaries, particularly in maintaining up-to-date lists of political actors, and spot-checking of the accuracy of the coding.

If these changes proceed on schedule, we are hoping that around September 2013 or thereabouts we should be in a position to do a complete GDELT 2.0 recode. In the meantime, however, we will first deploy the existing data in a mySQL database on a high-bandwidth server at the University of Texas at Dallas (with the assistance of Patrick Brandt) and implement the system for near-real-time coding with daily updates, which has been running in an experimental mode since October 2012. Updates on the status of these systems can be found at http://eventdata.psu.edu/data.dir/GDELT.html, which also contains a "reduced" version of the data set that has only the core event, actor and location variables (but for all countries and years). When the system is fully operational, we will broadcast announcements on various international relations and methodology listservs.

7 Appendix: GDELT CODEBOOK Version 1.0
Documentation Version Date: 25-July-2012

7.1 INTRODUCTION

This codebook provides a quick overview of the fields in the GDELT data file format and their descriptions. GDELT event records are in the dyadic CAMEO format, capturing two actors and the action performed by Actor1 upon Actor2. A wide array of variables break out the raw CAMEO actor codes into their respective fields to make it easier to interact with the data, the Action codes are broken out into their hierarchy, the Goldstein ranking score is provided, an average tone score is provided for all coverage of the
event, several indicators of importance are provided, and a special array of georeferencing fields offer estimated landmark-centroid-level geographic positioning of both actors and the location of the action.

A reduced version of the data set which contains the basic event, actor and geolocation variables for all countries for Jan-1979 to June-2012 can be downloaded from a link at http://eventdata.psu.edu/data.dir/GDELT.html. The file is about 650Mb compressed and includes Python programs for doing basic subsetting and generating counts and R graphics: the file \GDELT.reduced.documentation.txt" describes the file format and utility programs; files are in Unix format.

We are in the process of installing GDELT on a server at UT/Dallas with appropriate bandwidth; this installation will provide a mySQL facility for subsetting and automatically update the dataset on a daily basis.

7.2 BASE ATTRIBUTES

These attributes capture the date and raw actor codes for Actor1 and Actor2.

- **Day**. Date the event took place in YYYYMMDD format.
- **GlobalEventID**. Globally unique identifier assigned to each event record that uniquely identifies it in the master dataset.
  
  NOTE: While these will often be sequential with date, this is NOT always the case and this field should NOT be used to sort events by date: the date fields should be used for this.
- **MonthYear**. Alternative formatting of the event date, in YYYYMM format.
- **Year**. Alternative formatting of the event date, in YYYY format.
- **FractionDate**. Alternative formatting of the event date, computed as YYYY.FFFF, where FFFF is the percentage of the year completed by that day. This collapses the month and day into a fractional range from 0 to 0.9999, capturing the 365 days of the year. The fractional component (FFFF) is computed as (MONTH * 30 + DAY) / 365. This is an approximation and does not correctly take into account the differing numbers of days in each month or leap years, but offers a simple single-number sorting mechanism for applications that wish to estimate the rough temporal distance between dates.
- **Actor1Code**. The complete raw CAMEO code for Actor1 (includes geographic, class, ethnic, religious, and type classes). May be blank if the system was unable to identify an Actor1.
- **Actor1Name**. The actual name of the Actor 1. In the case of a political leader or organization, this will be the leaders formal name (GEORGE W BUSH, UNITED NATIONS), for a geographic match it will be either the country or capital/major city name (UNITED STATES / PARIS), and for ethnic, religious, and type matches it will reflect the root match class (KURD, CATHOLIC, POLICE OFFICER, etc). May be blank if the system was unable to identify an Actor1.
- **Actor2Code**. The complete raw CAMEO code for Actor2 (includes geographic, class, ethnic, religious, and type classes). May be blank if the system was unable to identify an Actor2.
- **Actor2Name**. The actual name of the Actor 2. In the case of a political leader or organization, this will be the leaders formal name (GEORGE W BUSH, UNITED NATIONS), for a geographic match it will be either the country or capital/major city name (UNITED STATES / PARIS), and for ethnic, religious, and type matches it will reflect the root match class (KURD, CATHOLIC, POLICE OFFICER, etc). May be blank if the system was unable to identify an Actor2.

7.3 ACTOR CODE BREAKOUT

The Actor1 and Actor2 fields may contain multiple codes indicating geographic, ethnic, and religious affiliation and the actors role in the environment (political elite, military officer, rebel, etc). These codes may be combined in any order, and are encoded in a single character field consisting of a string of concatenated 3-digit codes. To make it easier to utilize this information in analysis, this section breaks these codes out into a set of individual columns.

http://semanticommunity.info/AOL_Government/Managing_Big_Data_For_Results
Updated: Wed, 23 Sep 2015 07:38:19 GMT
Powered by mindtouch
<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor1CountryCode</td>
<td>The 3-digit CAMEO code for the country affiliation of Actor1.</td>
</tr>
<tr>
<td>Actor1CountryLabel</td>
<td>The human-readable name of the country affiliation of Actor1.</td>
</tr>
<tr>
<td>Actor1KnownGroupCode</td>
<td>If Actor1 is a known IGO/NGO/rebel organization (al-Qaeda, United Nations, World Bank, etc) with its own CAMEO code, this field will contain that code.</td>
</tr>
<tr>
<td>Actor1KnownGroupLabel</td>
<td>The human-readable formal name for Actor1KnownGroupCode.</td>
</tr>
<tr>
<td>Actor1EthnicCode</td>
<td>If the source document specifies the ethnic affiliation of Actor1 and that ethnic group has a CAMEO entry, the CAMEO code is entered here. NOTE: a few special groups like ARAB may also have entries in the type column due to legacy CAMEO behavior.</td>
</tr>
<tr>
<td>Actor1EthnicLabel</td>
<td>The human-readable formal name for Actor1EthnicCode.</td>
</tr>
<tr>
<td>Actor1Religion1Code</td>
<td>If the source document specifies the religious affiliation of Actor1 and that religious group has a CAMEO entry, the CAMEO code is entered here. NOTE: a few special groups like JEW may also have entries in the geographic or type columns due to legacy CAMEO behavior.</td>
</tr>
<tr>
<td>Actor1Religion1Label</td>
<td>The human-readable formal name for Actor1Religion1Code.</td>
</tr>
<tr>
<td>Actor1Religion2Code</td>
<td>If multiple religious codes are specified for Actor1, this contains the secondary code. Some religion entries automatically use two codes, such as Catholic, which invokes Christianity as Code1 and Catholicism as Code2.</td>
</tr>
<tr>
<td>Actor1Religion2Label</td>
<td>The human-readable formal name for Actor1Religion2Code.</td>
</tr>
<tr>
<td>Actor1Type1Code</td>
<td>The 3-digit CAMEO code of the CAMEO type or role of Actor1, if specified. This can be a specific role such as Police Forces, Government, Military, Political Opposition, Rebels, etc, a broad role class such as Education, Elites, Media, Refugees, or organizational classes like Non-Governmental Movement. Special codes such as Moderate and Radical may refer to the operational strategy of a group.</td>
</tr>
<tr>
<td>Actor1Type1Label</td>
<td>The human-readable formal name for Actor1Type1Code.</td>
</tr>
<tr>
<td>Actor1Type2Code</td>
<td>If multiple type/role codes are specified for Actor1, this returns the second code.</td>
</tr>
<tr>
<td>Actor1Type2Label</td>
<td>The human-readable formal name for Actor1Type2Code.</td>
</tr>
<tr>
<td>Actor1Type3Code</td>
<td>If multiple type/role codes are specified for Actor1, this returns the third code.</td>
</tr>
<tr>
<td>Actor1Type3Label</td>
<td>The human-readable formal name for Actor1Type3Code.</td>
</tr>
</tbody>
</table>

These codes are repeated for Actor2, using the prefix Actor2 instead of Actor1. As with Actor1, if no Actor2 could be extracted, these fields will be blank. Only in extremely rare circumstances will both Actor1 and Actor2 be blank.

### 7.4 EVENT ACTION ATTRIBUTES

These fields break out various attributes of the event action and offer several mechanisms for assessing the importance or immediate-term impact of an event.

- **IsRootEvent.** The system codes every event across an entire document, using an array of techniques to deference and link information together. A number of previous projects such as the ICEWS initiative have found that events occurring in the lead paragraph of a document tend to be the most important and are the least likely to have any errors. Thus, this flag can be used as a proxy for the rough importance of an event to create subsets of the event stream.

- **EventCode.** This is the raw CAMEO action code describing the action that Actor1 performed upon Actor2.

- **EventDesc.** This is the human-readable formal label for the given CAMEO action code.

- **EventBaseCode.** CAMEO event codes are dened in a three-level taxonomy. For events at level three in the taxonomy, this yields its level two leaf root node. For example, code 0251 (Appeal for easing of administrative sanctions) would yield
an EventBaseCode of 025 (Appeal to yield). This makes it possible to aggregate events at various resolutions of specificity. For events at levels two or one, this field will be set to EventCode.

- **EventBaseDesc.** This is the human-readable formal label for EventBaseCode.
- **EventRootCode.** Similar to EventBaseCode, this defines the root-level category the event code falls under. For example, code 0251 (Appeal for easing of administrative sanctions) has a root code of 02 (Appeal). This makes it possible to aggregate events at various resolutions of specificity. For events at levels two or one, this field will be set to EventCode.
- **EventRootDesc.** This is the human-readable formal label for EventBaseCode.
- **QuadClass.** The entire CAMEO event taxonomy is ultimately broken into four primary classifications: Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material Conflict. This eld species this primary classification for the event type, allowing analysis at the highest level of aggregation.
- **GoldsteinScale.** Each CAMEO event code is assigned a numeric score from -10 to +10, capturing the likely impact that type of event will have on the stability of a country. This is known as the Goldstein Scale. This field species the Goldstein score for each event type. NOTE that this score is based on the type of event, not the specifics of the actual event record being recorded thus two riots, one with 10 people and one with 10,000, will both receive the same Goldstein score. This can be aggregated to various levels of time resolution to yield an approximation of the stability of a geography over time.
- **NumMentions.** This is the total number of mentions of this event across all source documents. Multiple references to an event within a single document also contribute to this count. This can be used as a method of assessing the importance of an event: the more discussion of that event, the more likely it is to be significant. The total universe of source documents and the density of events within them vary over time, so it is recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest.
- **NumSources.** This is the total number of information sources containing one or more mentions of this event. This can be used as a method of assessing the importance of an event: the more discussion of that event, the more likely it is to be significant. The total universe of sources varies over time, so it is recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest.
- **NumArticles.** This is the total number of source documents containing one or more mentions of this event. This can be used as a method of assessing the importance of an event: the more discussion of that event, the more likely it is to be significant. The total universe of source documents varies over time, so it is recommended that this eld be normalized by the average or other measure of the universe of events during the time period of interest.
- **AvgTone.** This is the average tone of all documents containing one or more mentions of this event. The score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This can be used as a method of filtering the context of events as a subtle measure of the importance of an event and as a proxy for the impact of that event. For example, a riot event with a slightly negative average tone is likely to have been a minor occurrence, whereas if it had an extremely negative average tone, it suggests a far more serious occurrence. A riot with a positive score likely suggests a very minor occurrence described in the context of a more positive narrative (such as a report of an attack occurring in a discussion of improving conditions on the ground in a country and how the number of attacks per day has been greatly reduced).

### 7.5 EVENT GEOGRAPHY

The final set of fields add a novel enhancement to the CAMEO taxonomy, georeferencing each event along three primary dimensions to the landmark-centroid level. To do this, the fulltext of the source document is processed using fulltext geocoding and automatic disambiguation to identify every geographic reference. The closest reference to each of the two actors and to the action reference are then encoded in these fields. The georeferenced location for an actor may not always match the Actor1CountryCode eld, such as in a case where the President of Russia is visiting Washington, DC in the United States, in which case the Actor1CountryCode would contain the code for Russia, while the georeferencing fields below would contain a match for Washington.
DC. It may not always be possible for the system to locate a match for each actor or location, in which case one or more of the fields may be blank. Finally, the Action fields capture the location information closest to the point in the event description that contains the actual statement of action.

- **Actor1Geo Type.** This field specifies the geographic resolution of the match type and holds one of the following values: COUNTRY (the match was at the country level), USSTATE (the match was to a US state), USLOC (the match was to a US city or landmark), WORLDLOC (the match was to a city or landmark outside the US). This can be used to filter events by geographic specificity, for example, extracting only those events with a landmark-level geographic resolution for mapping. Note that both COUNTRY and USSTATE matches will still provide a latitude/longitude pair, which will be the centroid of that country or state. Matches to foreign Administrative Division 1s (ADM1s) (the rough equivalent of a US state) will be coded as a WORLDLOC location.
- **Actor1 Geo Fullname.** This is the full human-readable name of the matched location. In the case of a country it is simply the country name. For US states it is in the format of State, United States, while for all other matches it is in the format of Landmark, State/ADM1, Country. This can be used to label locations when placing events on a map.
- **Actor1Geo CountryCode.** This is the 2-character FIPS10-4 country code for the location.
- **Actor1Geo ADM1Code.** This is the 2-character FIPS10-4 administrative division 1 (ADM1) code for the administrative division housing the landmark. In the case of the United States, this is the 2-character shortform of the states name (such as TX for Texas).
- **Actor1Geo Lat.** This is the centroid latitude of the landmark for mapping.
- **Actor1Geo Long.** This is the centroid longitude of the landmark for mapping.

These codes are repeated for Actor2 and Action, using those prefixes.

### References


---

**Footnotes**

1

In view of recent developments in the U.S. Senate, we are placing more emphasis in this text on the degree to which this work has been possible due to funding from the NSF--predominantly the beleaguered NSF Political Science program--than might normally be the case. Thank you for your understanding; perhaps it will be contagious.

2

The complete GDELT data set incorporates additional stories from other sources.

3

[http://www.dlib.org/dlib/september12...09leetaru.html](http://www.dlib.org/dlib/september12...09leetaru.html)
Python and R code for generating these is included in the "GDEL.reduced" package that can be downloaded at http://eventdata.psu.edu/data.dir/GDEL.html.

Consult the internal documentation of these datasets for further information on search terms, filtering and so forth. As with GDELT, the Levant data use a "One-A-Day" filter, though unlike GDELT, this is based only on the dyad, not the event location.

We experimented with eliminating low frequency VERCP cases here and it did not make much difference.

An additional factor that might accounts for some of the differences is that fact that GDELT is using location-based duplicate filtering--events are considered duplicates only if they are the same event and occur in the same city--whereas as far as we know, ICEWS was using only dyad-based duplicate filtering.

"Not definitive": yes, that means you, BBN and MITRE. . .

There also appear to be some problems getting the C++ "ncurses" library, which Tabari uses for its interface, to run on some versions of Ubuntu Linux.

- See more at: http://gdelt.utdallas.edu/about.html...UKxgK5UT.dpuf

---

**Downloading GDELT**

Source: http://gdelt.utdallas.edu/data.html

There are three versions of the data available for download.

"Reduced" 1979-2012 Dataset

This version of the data contains only a subset of the data fields for each record and uses the "one-a-day" country-level filtering commonly used in previous country-level event datasets that did not have the city-level geographic resolution of GDELT. This version of the data will most closely match what users with previous event analysis experience are used to
working with in terms of aggregation level, but collapses the database on BASEATTRS+SOURCE+TARGET+EVENTCODE. This means that this version of the data often collapses multiple separate riots on the same day in different cities into a single record, making it much more difficult to trace city-level spatial patterns in unrest.

The Reduced dataset comes with a set of Python scripts that makes working with the data easier. Most of the existing tutorials have been built using this dataset and this version is the easiest when getting started with the data. However, it is lacking many of the extended fields that can be highly useful in filtering the data, such as the geographic resolution field and does not break out the Actor codes, requiring keyword-based searches to look for specific actor types.

If you are looking for a quick way to start experimenting with GDELT and get your feet wet, this is a good place to start, especially since it lets you explore patterns over nearly a quarter-century from 1979 to 2012. However, if you are planning to work extensively with GDELT for publication or to use the daily event stream for watchboarding, forecasting, or other work, you may want to just go ahead and start with Daily Updates files so that you have access to all of the extended event record fields and can develop your tools to work with this version of the data, instead of working with the Reduced version and then having to adjust all of your tools for the different format of the Historical and Daily Updates files.

Note that this version of the data differs substantially from the Histroical and Daily Updates files, and their codebook is not applicable here.

**Download**

- [Download Reduced January 1979 - June 2012 Dataset](#) (667MB) My Note: I downloaded this.

---

**Historical 1979 - March 2013 Backfile**

This is the full resolution GDELT dataset running January 1, 1979 through March 31, 2013 and containing all data fields for each event record. The years 1979 through 2005, inclusive, are available as yearly downloads containing all records for each year, while starting in January 2006 data is available as monthly downloads due to the larger number of records per month over time.

**Download**

- [Download Historical Backfiles](#)
- [Column Labels Header Row - Historical Backfiles](#)
- [Column Labels With Their Column IDs](#) (Quick reference for parsing the fields in a scripting or programming environment.)
- [GDELT Data Format Codebook/Documentation](#)
- [CAMEO Code Reference](#)

---

**Daily Updates April 1, 2013 to Present**

This is the full resolution GDELT dataset containing all data fields for each event record and updated daily beginning April 1, 2013. It contains an additional field at the end of each record that is not in the backfiles, which is the Source URL, which gives the Source URL of the article the event was found in.
Each morning, seven days a week, the latest daily update is posted by 4AM CST. This file is named with the previous day's date in the format "YYYYMMDD.export.CSV.zip" (ie the morning of May 24, 2013 a new file called "20130524.export.CSV.zip" is added). UNIX or Linux users can easily set up a cronjob or other automatic scheduling processes to automatically download the latest daily update each morning and process it for watchboarding, forecasting, early warning, alert services, and other applications.

**NOTE** that while each of the Backfile ZIPS record only those events occurring during the specific month/year of the file, the daily files contain all events found in news coverage published on that day. This means that the vast majority of each daily file will be events that occurred that day, but there will also be a certain percentage of events from previous days that were mentioned in coverage of that day. Thus, users should perform basic date filtering to ensure that only events during the desired time frame(s) are considered.

**Download**

- [Download Daily Update Files](#)
- [Column Labels Header Row - Daily Updates](#)
- [Column Labels With Their Column IDs](#) (Quick reference for parsing the fields in a scripting or programming environment.)
- [Sample Daily Downloader PERL Script](#) (You can set this script up via cronjob on a Linux computer to automatically download the latest daily update file each morning.)
- [GDELT Data Format Codebook/Documentation](#)
- [CAMEO Code Reference](#)

---

**CAMEO Lookups and SQL Code for Historical and Daily Updates**

The data format of the Historical and Daily Updates versions of GDELT record the raw CAMEO 3-digit actor codes and numeric event codes. These lookups give the textual labels for each of those fields making it easier to work with the data for those who have not previously worked with CAMEO.

**CAMEO Actor Code Lookups**

- [CAMEO Lookup - Country Codes](#)
- [CAMEO Lookup - Known Group Codes](#)
- [CAMEO Lookup - Ethnic Codes](#)
- [CAMEO Lookup - Religion Codes](#)
- [CAMEO Lookup - Type Codes](#)

**CAMEO Event Code Lookups**

- [CAMEO Lookup - Event Codes](#)
- [CAMEO Lookup - Goldstein Scale](#)

Advanced users will likely wish to store the GDELT data in a relational database system to make it easier to interact with the complete dataset using SQL. The following links provide a sample SQL schema for the table, sample loading SQL statements, and sample export SQL statements.
SQL Bulk Loader Code

- [SQL Table Definition](#) (Use this to create a table that you can bulk-import GDELT into).
- [SQL Bulk Loader Example](#) (Use this to bulk-import GDELT into the tables above).
- [SQL Bulk Exporter Example](#) (Use this to bulk-export a subset of GDELT data from the above tables for analysis).

FAQ

Source: [http://gdelt.utdallas.edu/faq.html](http://gdelt.utdallas.edu/faq.html)

**What sources does GDELT rely upon?**

GDELT is based purely on unclassified mainstream news media reports from throughout the world. See the [About](#) page for more detail.

We are actively exploring incorporating a range of social media sources, but work is ongoing there.

**Can you send me the source news articles behind your event database?**

Due to copyright restrictions and publisher agreements we cannot redistribute any of the news content that was used in the creation of GDELT, only the codified numeric event records extracted from that content. However, for content after April 1, 2013 we do include the URL or citation to the source article for each event so that you can locate the material on your own to read more about the event and its surrounding context. In the next release of GDELT, tentatively slated for late Fall 2013, we will be including source citations for all events back to 1979.

In addition, the Fall 2013 release of GDELT 2.0 will include a vast array of new codified emotional and thematic indicators such as "View Towards Government", "Focus on Healthcare", "Emphasis on Women's Rights", etc. Each of these will offer a numeric score codifying, by city, the attention and emotional response towards each of these dimensions by day, allowing for rich contextualization of the GDELT physical event database.

**How does GDELT compare with the US DOD ICEWS?**

GDELT is extremely similar to ICEWS in concept and indeed uses many of the same core underpinnings, but includes a vast array of enhancements over ICEWS and uniquely is fully unclassified and available for open academic and commercial research, updated daily.

**Key Differentiators**

- GDELT includes a **multi-stage filtering and linguistic rewriting pipeline** that significantly increases TABARI's accuracy
- GDELT uses a **sophisticated full-text disambiguating geocoding system**, one of just a handful of production systems available, which offers massively **enhanced geographic recovery and resolution**
• GDELT assigns events to the **city or landmark-level**, while ICEWS is limited to a maximum resolution of an administrative division (roughly equivalent to a US state) - with ICEWS this is equivalent to stating that a riot took place "somewhere in the State of California", while GDELT gives you the **specific city**

• GDELT also captures the **city-level** geographic affiliation of both of the actors involved in the event, allowing for **city-level tracing of leader movements and diplomatic activity**

• GDELT provides more **sophisticated matching** under the ambiguous reporting common to **high-conflict zones** such as Syria and Afghanistan

• GDELT captures both **religion and ethnic group membership** information where available

• GDELT covers all **countries in the world** over more than three decades

• GDELT covers the entire past quarter-century: **1979 to present**

• GDELT is **100% open, unclassified, and available for unlimited use and redistribution**

---

**How does GDELT compare with other event databases?**

Leetaru & Schrodt’s (2013) International Studies Association paper announcing the data includes several comparisons to existing datasets and many more comparisons are actively underway. GDELT currently is the only database to cover nearly all countries globally back to 1979 with all events georeferenced and daily updates available. See the About page for a list of key features that make GDELT unique.

---

**I've previously used TABARI for a coding project and have an event database, how can I contribute?**

Please contact us, as we’re currently exploring setting up a community contribution program where users like you can contribute your event databases (with full credit and citation to you) to offer expanded coverage of specific geographic regions, social or insurgent groups, and topics.

---

**I have some ideas for additional event categories you should add.**

We’d love to hear from you! In particular, we’re looking for help in developing rigorous new taxonomies, such as for human rights violations, expanding the underlying grammars to recognize new event types, and finding sufficient examples of each event to be able to robustly test the accuracy of the system.

---

**I’m a computer science researcher with some NLP tools that might be useful.**

We’re extremely interested in hearing from computer scientists and other researchers working in the areas of Natural Language Processing and other areas of research that are developing tools that could help us in our quest to constantly enhance and upgrade the event extraction core of GDELT. There are already several efforts underway developing completely new event extraction algorithms and pipelines to replace TABARI, which we hope to roll out in GDELT 2.0’s release this fall.

GDELT relies heavily on many areas of active research in the NLP community, including entity recognition, relationship extraction, fact and claim extraction, separating potential from actual statements, pronoun coreference resolution, fulltext geographic disambiguation, and a vast array of other areas, and we need tools that are robust against OCR’d text.
Managing Big Data For Results

Source: http://fcw.com/events/bigdata/home.aspx

Big Data: A Revolution That Will Transform How We Live, Work, and Think

Source: http://www.amazon.com/books/dp/0544002695

Viktor Mayer-Schonberger & Kenneth Cukier


Book Description

Release date: March 5, 2013

A revelatory exploration of the hottest trend in technology and the dramatic impact it will have on the economy, science, and society at large.

Which paint color is most likely to tell you that a used car is in good shape? How can officials identify the most dangerous New York City manholes before they explode? And how did Google searches predict the spread of the H1N1 flu outbreak?

The key to answering these questions, and many more, is big data. "Big data" refers to our burgeoning ability to crunch vast collections of information, analyze it instantly, and draw sometimes profoundly surprising conclusions from it. This emerging science can translate myriad phenomena—from the price of airline tickets to the text of millions of books—into searchable form, and uses our increasing computing power to unearth epiphanies that we never could have seen before. A revolution on par with the Internet or perhaps even the printing press, big data will change the way we think about business, health, politics, education, and innovation in the years to come. It also poses fresh threats, from the
inevitable end of privacy as we know it to the prospect of being penalized for things we haven’t even done yet, based on big data’s ability to predict our future behavior.

In this brilliantly clear, often surprising work, two leading experts explain what big data is, how it will change our lives, and what we can do to protect ourselves from its hazards. Big Data is the first big book about the next big thing.

Editorial Reviews

Amazon.com Review

Amazon Exclusive: Q&A with Kenneth Cukier and Viktor Mayer-Schonberger

Q. What did it take to write Big Data?

A. Kenn has written about technology and business from Europe, Asia, and the US for The Economist, and is well-connected to the data community. Viktor had researched the information economy as a professor at Harvard and now at Oxford, and his book Delete had been well received. So we thought we had a good basis to make a contribution in the
area. As we wrote the book, we had to dig deep to find unheard stories about big data pioneers and interview them. We wanted *Big Data* to be about a big idea, but also to be full of examples and success stories -- and be engrossing to read.

**Q. Are you big data’s cheerleaders?**

A. Absolutely not. We are the messengers of big data, not its evangelists. The big data age is happening, and in the book we take a look at the drivers, and big data’s likely trajectory: how it will change how we work and live. We emphasize that the fundamental shift is not in the machines that calculate data, but in the data itself and how we use it.

**Q. In discovering big data applications, what was your biggest surprise?**

A. It is tempting to say that it was predicting exploding manholes, tracking inflation in real time, or how big data saves the lives of premature babies. But the biggest surprise for us perhaps was the very diversity of the uses of big data, and how it already is changing people’s everyday world. Many people see big data through the lens of the Internet economy, since Google and Facebook have so much data. But that misses the point: big data is everywhere.

**Q. Is Big Data then primarily a story about economic efficiency?**

A. Big data improves economic efficiency, but that’s only a very small part of the story. We realized when talking to dozens and dozens of big data pioneers that it improves health care, advances better education, and helps predict societal change—from urban sprawl to the spread of the flu. Big data is roaring through all sectors of the economy and all areas of life.

**Q. So big data offers only “upside”?**

A. Not at all. We are very concerned about what we call in our book “the dark side of big data.” However the real challenge is that the problem is not necessarily where we initially tend to think it is, such as surveillance and privacy. After looking into the potential misuses of big data, we became much more troubled by “propensity” -- that is, big data predictions being used to police and punish. And by the “fetishization” of data that may occur, whereby organizations may blindly defer to what the data says without understanding its limitations.

**Q. What can we do about this “dark side”?**

A. Knowing about it is the first step. We thought hard to suggest concrete steps that can be taken to minimize and mitigate big data’s risk, and came up with a few ways to ensure transparency, guarantee human free will, and strike a better balance on privacy and the use of personal information. These are deeply serious issues. If we do not take action soon, it might be too late.

**Reviews**

"Every decade, there are a handful of books that change the way you look at everything. This is one of those books. Society has begun to reckon the change that big data will bring. This book is an incredibly important start."

—Lawrence Lessig, Roy L. Furman Professor of Law, Harvard Law School, and author of Remix and Free Culture

"This brilliant book cuts through the mystery and the hype surrounding big data. A must-read for anyone in business, information technology, public policy, intelligence, and medicine. And
anyone else who is just plain curious about the future."
—John Seely Brown, former Chief Scientist, Xerox Corp., and head of Xerox Palo Alto Research Center

"Big Data breaks new ground in identifying how today’s avalanche of information fundamentally shifts our basic understanding of the world. Argued boldly and written beautifully, the book clearly shows how companies can unlock value, how policymakers need to be on guard, and how everyone’s cognitive models need to change."
—Joi Ito, Director of the MIT Media Lab

"Big Data is a must-read for anyone who wants to stay ahead of one of the key trends defining the future of business."
—Marc Benioff, Chairman and CEO, salesforce.com

"An optimistic and practical look at the Big Data revolution — just the thing to get your head around the big changes already underway and the bigger changes to come."
—Cory Doctorow, boingboing.com

"Just as water is wet in a way that individual water molecules aren’t, big data can reveal information in a way that individual bits of data can’t. The authors show us the surprising ways that enormous, complex, and messy collections of data can be used to predict everything from shopping patterns to flu outbreaks."
—Clay Shirky, author of Cognitive Surplus and Here Comes Everybody

"The book teems with great insights on the new ways of harnessing information, and offers a convincing vision of the future. It is essential reading for anyone who uses — or is affected by — big data."
—Jeff Jonas, IBM Fellow & Chief Scientist, IBM Entity Analytics

“What I’m certain about is that Big Data will be the defining text in the discussion for some time to come."
—Forbes.com

“The authors make clear that ‘big data’ is much more than a Silicon Valley buzzword… No other book offers such an accessible and balanced tour of the many benefits and downsides of our continuing infatuation with data.”
—Wall Street Journal

"Plenty of books extol the technical marvels of our information society, but this is an original analysis of the information itself—trillions of searches, calls, clicks, queries and purchases….A fascinating, enthusiastic view of the possibilities of vast computer correlations and the entrepreneurs who are taking advantage of them."
—STARRED Kirkus Reviews

"This book offers important insights and information"
—Booklist

"'big data' [is] one of the buzzwords of corporate executives, tech-savvy politicians, and worried civil libertarians. If you want to know what they’re all talking about, then Big Data is the book for you, a comprehensive and entertaining introduction to a very large topic….Mayer-Schönberger and Cukier offer up some sensible suggestions on how we can have the blessings of big data and our freedoms, too. Just as well;
The precise definition of what constitutes big data does not exist, it is a term used to refer to the capture of enormous amounts of different types of data that often seems to be unrelated. Yet, that imprecise definition is part of the strength of using big data to make better decisions. In the days when only small samples could be taken for analysis due to the cost, it was critical that everything be done right, the items in the sample must be randomly chosen and care had to be taken to eliminate any extreme outliers that would skew the result. This also meant that the models had to be very well constructed, for if the model was not applicable, the final results could be worthless or even have negative consequences.

The concept of big data basically means that all the data is examined to look for common characteristics. Outliers are included and are of less significance for they will be drowned out by the enormous number of data points in the middle. One of the examples of the use of big data is the prediction of high fevers in infants. Rather than developing a model for the events that would include many assumptions, not all of which are correct, the immediate history of the children that develop high fevers is examined. All of the vital signs and other data collected about the infants are then examined to determine if there are any common indicators that could be used as predictors. The data analysts are not trying to establish causality, only traits present before the events.

Doing this means that only the data matters, emotion and experience are almost insignificant. The authors describe many examples of where big data has been used to predict and prioritize; one of the most interesting examples is the development of translation software. Rather than use a team of translators to develop the conversion rules, an enormous number of documents that have already been translated from one language to another are examined and used to build the models used in translation. This has been so effective that there is a joke that the efficiency of the translation software is greater when the linguists are not involved.

There is an enormous amount of potential value in the examination of big data; some of the successes so far have been surprising, yet understandable in retrospect. My favorite was Wal Mart discovering that when there are predictions of a hurricane, stores in the projected area to be hit sell a lot of Pop Tarts. As a teacher of statistics and occasional consultant, I was a bit surprised to learn that the big data revolution does not involve a lot of statistics; in fact statistical thinking is discouraged by many that work in big data. The goal of big data analysis is to find correlations, not causes.

If you have an interest in the mile-high view of what big data is and how it is being applied, this is an excellent book to satisfy that curiosity. In a world where the demographics and economic trends seem to doom us to a cycle of extreme austerity, big data has the potential to provide a great deal of relief. Learning how it is being used can help make you more optimistic about the future.
search terms that, appropriately weighted, strongly correlated with official data. However, unlike the CDC, Google was able to make those assessments in real time, not a week or two later.

Oren Etzioni, frustrated to learn that many passengers booking a flight after he had, were able to pay less - contrary to conventional wisdom. He then 'scraped' information from a travel website from a 41-day period to forecast whether a price was a good deal or not, founding Farecast to offer this new ability. Etzioni next went on to improve the system by digesting data from a travel site that covered most American commercial routes for a year - nearly 200 billion flight-price records. Before expanding to hotel rooms, concert tickets and used cars, Microsoft snapped up his firm ($110 million) and incorporated it into it Bing.

New processing technologies like open-source Hadoop allow managing far larger quantities of data. Hadoop uses a computational paradigm named MapReduce (by Google) to divide an application into many small fragments, each of which may be executed on any computer node in a cluster. Visa was able to reduce processing time for two years worth of data (73 billion transactions) from 1 month to 13 minutes using Hadoop.

The authors define 'big data' as things that can be done on a large scale that cannot be done on a smaller one, and see it as offering a major transformation. Potential sources include that approximately 7 billion shares change hands every day in finance, two-thirds via computer model direction, Google processing over 24 petabytes of data/day, Facebook getting over 10 million new photos uploaded every hour.

Random sampling was an earlier data analysis breakthrough, but it doesn't work well if users want to make smaller and smaller subgroup analyses. It no longer makes much sense when we can harness large amounts of data. Detection of credit card fraud works by looking for anomalies, and the best way to find them is to crunch all the data, even at an individual level - definitely a big data problem.

Xoom specializes in the analysis of international money transfers - it raised an alarm in 2011 after noticing a slight increase in Discover Card transactions from N.J. - they came from a criminal group.

Teaching computers to translate requires not only teaching them rules, but the exceptions as well. IBM researchers in 1980 improved the current state of the art using statistical probabilities derived from 3 million sentences pairs, but the results still weren't good enough. Google tried again in 2006 using every translation it could find - eg. corporate websites in multiple languages, translations of official documents, and translations of books from Google's book-scanning project. They compiled 95 billion English sentences and translations, some of dubious quality. It offers translations of over 60 languages and the results are more accurate than others, though still highly imperfect. The improvement came from using more data.

PriceStats uses a software crawler to collect a half-million prices of products sold in the U.S., daily. Analysis then allows immediate detection of price change trends, vs. BLS processes that take about two weeks to complete and cost about $250 million/year.

A third of Amazon's sales reportedly result from its correlation-based recommendation system - putting local booksellers at a decided disadvantage. Netflix has a similar system to boost its rental volume.

Bottom-Line: 'Big Data' was more up-to-date than prior works on the subject, but provided very little in explanation of the specifics of new uses. I was especially disappointed that it had failed to report on new applications in health care that would improve the quality and cost-effectiveness of care.

6 of 6 people found the following review helpful

5.0 out of 5 stars **Transformative!!** May 13, 2013
By litaddiction VINE™ VOICE

"Big data is about three major shifts of mindset that are interlinked and hence reinforce one another. The first is the ability to analyze vast amounts of data about a topic rather than be forced to settle for smaller sets. The second is a
willingness to embrace data's real-world messiness rather than privilege exactitude. The third is a growing respect for correlations rather than a continuing quest for elusive causality."

That's "big data" the *concept*, to which my reactions were, respectively, bogglement, disagreement, and suspicion. And then there's BIG DATA the *book*, wherein the authors unpacked their ideas and totally transformed mine.

First, about the mind-numbing amount of data, gathered everywhere -- Google and Facebook and public surveillance cameras for sure, but suffice it to say that everything electronic is gathering data, and everything that connects to the Internet is uploading the data to someone. And about the format of data, which has morphed from 75% *analog* in 2000 to 93% *digital* in 2007 (estimated to be >98% in 2013). Second, that the tidy, structured data of relational databases is now miniscule (estimated at 5%) compared with the as-yet untapped, error-ridden stuff of real life, like blogs and video. And third, that conceiving hypotheses, gathering perfect, representative data, and reaching causal conclusions is nowhere near as valuable or timely as finding correlations (the "what, not why") in a gigantic mess of data. The authors characterize big data as, "the equivalent of impressionist painting, wherein each stroke is messy when examined up close, but by stepping back one can see a majestic picture." Fascinating!

Then they address the problems of big data and, unlike most "alarmist" book I've read, they propose solutions. They advise that the ship has sailed on individuals being in control of their private information and online footprints (e.g. via opting out or being anonymous), especially with the secondary and tertiary (and quaternary, and...) markets that re-analyze data long after it's been collected. So they suggest that the data *users* be held accountable through law/ regulation similar to what's in place for other industries that hold potential for public harm. They suggest a new professional -- a "data scientist" or "algorithmist" -- who isn't the do-er who queries big data but rather the outside-the-lines thinker with a big-data mindset who "peers into databases to make a discovery" that creates new value. And they caution against "what's-past-is-prologue" thinking -- where personal history and the statistics of correlation drive everything from basing your credit score upon the credit scores of your Facebook friends, to *Minority Report*-like "predictive policing" -- arguing instead for safeguards that recognize free will and actual behaviors.

Here is a book with the *awe* I've been seeking! I turned every page with excitement about what would be on the next page. There's some repetition, but it's usually with a twist that enhances internalization and recollection, and there are dozens of fascinating business examples along the way. It's optimistic not alarmist; rather than running to find a doomsday hidey-hole, I came away transformed. It's the best book I've read so far this year.
5 of 5 people found the following review helpful

3.0 out of 5 stars An O.K. read, but nothing special April 30, 2013
By John Haldi

The discussion ranges far and wide, all over the map. Many, many things are touched upon, virtually none in depth. Lots of anecdotes.

10 of 12 people found the following review helpful

3.0 out of 5 stars Interesting Ideas, a Little Light on Substance April 9, 2013
By ironman96 VINE™ VOICE

First, I agree with the authors' premise that "Big Data" is going to transform the world in ways we don't even understand yet. The book did a decent job of explaining "Big Data" and gave a few examples. What left me somewhat disappointed was a paucity of real examples and constant repetition. It was as if the book was actually an expansion of a long magazine article--forcing the authors to repeat much of the same content over and over to provide enough filler for a book. There were some very intriguing examples of real world applications and suggestions of future uses of data. However, the book would have been better with more real world examples. Now maybe we're a little early into this and there are fewer publicly available examples. Regardless, the book promised a lot and then didn't deliver to the level I expected. I'm still glad to have read the book and gotten a better understanding of "Big Data" as I believe the authors are right about the changes it is bringing--some good, and some not so good.

3 of 3 people found the following review helpful

4.0 out of 5 stars Excellent high level book about Big Data and the struggles that occur May 10, 2013
By Daniel J Johnson

For anyone looking for a great primer on Big Data and the concerns that surround it, this is the book for you. I would highly recommend for business analysts and managers, including c-level execs. Mayer-Schonberger does a great job on identifying the key issues around Big Data and offering his opinion and insights on how we should move forward.

The only reason why I did not give this a five-star review is that the beginning starts off a bit slow and then the book hits it's stride about midway through. Truthfully, this would be a solid 4 1/2 star review, if Amazon allowed. If you can be patient through the first few chapters, you will not be disappointed. However, if you are completely new to the Big Data revolution, this books would make my top five list of must-reads to get your mind around the phenomenon. The first few chapters do a great job setting the stage.

To the initiated in Big Data, there are some fantastic arguments and well thought out opinions on how the industry should proceed as a whole. Frankly, I know I am wiser and have a more rounded understanding after reading. Should make any Big Data person's bookshelf.

(Just in case the author ever reads this review - I appreciate that you wrote this book for a broad audience. I would love to read some material written by you that is more focused on the issues surrounding Big Data. As an example, I think you could do a great job of writing a book simply on the ethics of data and our responsibilities as data stewards)

3 of 3 people found the following review helpful

4.0 out of 5 stars Big data renders the comprehension of the cosmos and the proof of the existence of God irrelevant May 9, 2013
By Brian Kodi VINE™ VOICE
Well, sorta. Big data marginalizes statistical sampling. With advances in technology, we are able to store and query comprehensive information such that the statistical sample or "N" will encompass the entire population. And the authors spend too many pages to drill this point into the reader's head, in my opinion. While the breadth of the information processed is approaching limitless, Big Data prediction is not exact, nor is it perfectly accurate. It facilitates predictions of high probabilities using correlations. Big data does not explain causation. When certain conditions are detected, specific outcomes result, but we may not know why. One of the more intriguing applications of Big Data is explained in chapter 4:

IBM and a team of university researchers have devised a software that foretells preemie (premature baby) infections far in advance. Early detection allows doctors to use lighter medical interventions to cure the ill. This was made possible by detecting a counterintuitive occurrence in analyzing 1,260 data points a second on the preemies' vital signs. When the vital signs are "very constant", serious infections often follow. No one knows why, but analysis of large data uncovered this hidden association.

As Big Data predictions become more accurate, we may reach a point of punishing people for acts that have yet to occur, much like in the movie "Minority Report". There are numerous other pitfalls of this advancing technology, and chief among them is data privacy rules that would have to be rewritten.

By and large, this book is very informative, and seems to have covered the good, the bad and the ugly of harnessing the power of Big Data. There are long stretches in the book that may seem bland to some readers, but the message - and the authors perceive themselves as messengers, not Big Data evangelists - is too powerful to be ignored. Well worth the read.

"Big Data" is not a new concept...

Source: http://fcw.com/events/bigdata/information/conference-overview.aspx

...for most agencies, but the opportunity to derive valuable and actionable information from analyzing and integrating the data collected by these organizations has expanded exponentially thanks to today’s technologies. From the data center to the desktop, government professionals increasingly are able to integrate and analyze the volume and variety of the enterprise data amass—and few enterprises collect more information than the combined Federal agencies. This FCW Executive Briefing will focus on tools, technologies and management strategies being developed to help government organizations manage and use their data to gain insights, solve problems, spur innovation, and discover valuable business intelligence contained in the big data they are receiving.

Attendees will learn:

• A overview of big data today and its role in agency decision-support moving ahead
• How to evaluate available government information assets and how to use what your organization already has on hand
• What tools are available for data mining and secure access to aggregated data
• How to engage in collaborative efforts across organizational and agency boundaries by unlocking the power of stove-piped data resources
• Benefits of access to diverse, immense data networks and how to protect them
• Real-world initiatives underway to exploit big data for practical, mission-driven purposes
• A survey of analytical and technological tools available to support agency programs to integrate evolving information collection and processing strategies
Agenda

Source: http://fcw.com/events/bigdata/information/agenda.aspx

7:30 - 8:20AM Registration, Continental Breakfast & Networking

8:20 - 8:25AM Welcome & Event Overview
Steve LeSueur [Executive Host]
Contributing Editor, 1105 Government Information Group

8:25 - 9:05AM Opening Keynote
Applying Analytics to Detect Fraud and Ensure Compliance
C. Scot Atkins
Program Manager, Supercomputing and Revenue Protection, U.S. Postal Service

9:05 - 9:45AM Session 2
How Big Data Analysis Supports Scientific and Medical Research Programs
Dr. Jack Collins
Director, Advanced Biomedical Computing Center, National Cancer Institute at Frederick, National Institutes of Health, Department of Health and Human Services

9:45 - 10:00AM Networking Break and Visit Exhibits

10:00 - 10:40AM Session 3
Keys to Effective Information Access—Computation, Correlation, and Collaboration
Niall Brennan
Director, Office of Information Products and Data Analytics, Office of Enterprise Management, Centers for Medicare and Medicaid Services

10:40 - 10:45AM Transition Break

10:45 - 11:15AM Session 4A
Data to Decision: What Your Data Can Do For Your Agency
Patrick Sims
Senior Director, Global Center of Excellence for Analytics, SAP

Session 4B
Control, Manage, and Analyze Big Data—Practical Approaches
Dave Denson
Big Data Solution Architect, NetApp

11:20 - 11:50AM Session 5A
Using Agency Big Data to Transform Operations
Gary Newgaard

Session 5B
Mastering the Art and Science of Big Data Management
Bob Kimball
11:55 - 12:25PM
Session 6A
How Big Data Applications Support Streamlined Operations and Risk Management
Slides and Paper
Clark Richey
Technical Director, Public Sector, MarkLogic

12:25 - 1:30PM
Luncheon and Plenary Session
Data Analytics in Action Slides
Chip Copper, Ph.D.
Principal Engineer, Brocade
Suzi Iacono, Ph.D.
Deputy Assistant Director, Computer and Information Science and Engineering, National Science Foundation
Marion Royal
Program Director, Program Management Office, Data.Gov, General Services Administration
Robert Ames [Moderator]
Senior Vice President for Information and Communications Technology, In-Q-Tel

1:40 - 2:00PM
Closing Remarks
Steve LeSueur [Executive Host]
Contributing Editor, 1105 Government Information Group
Visit Exhibits and Networking

Speakers

Source: http://fcw.com/events/bigdata/information/speakers.aspx

Government Speakers

C. Scot Atkins

Program Manager, Supercomputing and Revenue Protection
U.S. Postal Service

http://semanticommunity.info/AOL_Government/Managing_Big_Data_For_Results
Updated: Wed, 23 Sep 2015 07:38:19 GMT
Powered by mindtouch™
Scot Atkins is a Program Manager for the US Postal Service, holds a degree in mathematics and physics, and architects high density supercomputing (HDSC) systems providing world class operational and revenue protection solutions to real-time, big data challenges pushing the limits of volume, velocity, and variation (3V’s). Scot has been with the USPS for 20 years as a Program Manager, Operations Research Analyst, and acting Industrial Engineer, and previously served with the US Army as a Special Operations R&D Pilot and Test Officer for 10 years.

Supercomputing and Special Operations have represented the cutting edge in requirements challenges and creative innovation while Revenue Protection and Special Operations have represented the challenge of protecting against persistent threat. Scot has combined these nearly 30 years of experience to provide state-of-the-art IT architectural solutions to what may arguably be one of the world’s most voluminous real-time transactional data problems, while dispelling and demystifying the mutually exclusive myth of modern high density supercomputing vs. cloud computing.

Niall Brennan
Director, Office of Information Products and Data Analytics
Office of Enterprise Management, Centers for Medicare and Medicaid Services

Niall Brennan is the Director of the Office of Information Products and Data Analytics at the Centers for Medicare and Medicaid Services where he oversees agency efforts on data analytics and dissemination, and information products. The goal of the Office is to maximize the utility and availability of CMS data for internal and external users. Specifically, the Office performs a variety of analyses on high profile issues on a quick turnaround basis, for the CMS administrator, senior leaders and other HHS department leaders. Additionally, the Office is also responsible for coordinating the agency’s strategies regarding data storage, analysis and dissemination.

Prior to leading OIPDA, Brennan was Deputy Director and Acting Director of the Policy and Data Analysis Group at CMS. In this capacity he directed the evolution of more advanced agency analytics and led CMS participation in many transparency initiatives, including the Health Data Initiative. Additionally, Niall led the rulemaking and implementation of Sections 6002 and 10332 of the Affordable Care Act.

Prior to joining CMS, Brennan worked at the Brookings Institution where he directed a range of efforts to better measure both the quality and cost of health care. Previously, Brennan was a Senior Analyst at the Medicare Payment Advisory Commission (MedPAC) where he directed studies on measuring physician quality and cost, and the Medicare
Advantage program, and a Principal Analyst in the Budget Analysis Division at the Congressional Budget Office, where he worked on estimates related to Medicare reform, the Medicare drug benefit, and the Medicare hospital outpatient prospective payment system. He has also worked at The Urban Institute, and Price WaterhouseCoopers.

Brennan is a graduate of University College Dublin, and he earned a master's degree in Public Policy from Georgetown University.

**Jack Collins, Ph.D.**

Director, Advanced Biomedical Computing Center, National Cancer Institute at Frederick, National Institutes of Health
Department of Health and Human Services

Dr. Collins is the director of the Advanced Biomedical Computing Center at the Frederick National Laboratory for Cancer Research. Dr. Collins' research focuses on biomedical computing applications pertaining to cancer. His research group develops and applies high-performance algorithms to solve data-intensive computational biology problems in the areas of genomic analysis, pattern recognition in proteomics and imaging, molecular modeling, and systems biology.

He received his BA degree in mathematics from Park College with minors in physics, chemistry, and economics. His Ph.D. is in physical chemistry (theoretical chemistry) from the University of Nebraska at Lincoln.

Prior to joining NCI, Dr. Collins worked at SRI International (1985 to 1989) as a research associate and the Molecular Research Institute (1989 to 1992) as director of Computational Biology. Dr. Collins has over 70 peer-reviewed scientific publications and 3 patents. He has given dozens of invited talks and Webinars at international and national meetings and conferences, and serves on numerous review panels and advisory committees.

**Suzi Iacono, Ph.D.**

Deputy Assistant Director, Computer and Information Science and Engineering
National Science Foundation

Dr. Suzi Iacono is currently the Deputy Assistant Director for Computer and Information Science and Engineering (CISE) at the National Science Foundation (NSF). Among other responsibilities, she serves as the co-chair of the Big Data...
Senior Steering Group, the focal point for USG interagency Big Data activities under the Networking Information Technology Research and Development (NITRD) program. She also provides scientific leadership in the development of a federated advanced network testbed now being deployed across US campuses under the Global Environment for Network Innovations (GENI) program.

Over the past fourteen years at NSF, she has served in many capacities, including Division Director (DD) and Deputy DD for Information and Intelligent Systems, DD for the Division of Computer and Network Systems (CNS), and Acting Deputy of CISE. She headed up the Information Technology Research (ITR) Program, an NSF-wide Priority Area, and was the Program Director for Digital Society and Technologies.

Prior to coming to NSF, she held a faculty position at Boston University, was a Visiting Scholar at the Sloan School, Massachusetts Institute of Technology, and was a Research Associate at the Public Policy Research Office at the University of California, Irvine.

Over the years, she has written journal articles, book chapters and conference papers on Social Informatics, an area of interdisciplinary research and education that integrates aspects of computer and social sciences. Suzi received her PhD from the University of Arizona in Information Systems and her MA and BA from the University of California, Irvine in Social Ecology.

Marion Royal

Program Director, Program Management Office, Data.Gov
General Services Administration

Marion A. Royal is a nationally-recognized and Agency spokesperson on a diverse range of issues related to Electronic Government with specific emphasis on Enterprise Architecture, Data Interoperability and Collaboration. Royal is currently serving as Program Director on Data.gov. As a priority Open Government Initiative for President Obama's administration, Data.gov increases the ability of the public to easily find, download, and use datasets that are generated and held by the Federal Government. Data.gov provides descriptions of the Federal datasets (metadata), information about how to access the datasets, and tools that leverage government datasets.

Royal has worked on a number of governmentwide projects during his 17-year tenure at General Services Administration. Working with Federal Agencies, commercial organizations, and industry forums, he has served as a catalyst to advance International Data Harmonization, Governmentwide Electronic Messaging, Electronic Directories (X.500), Public Key Infrastructures, and Information Management and XML. He also served as Chief Architect during the Quicksilver task force that launched significant e-gov initiatives and he remains active with the Architecture and Infrastructure Committee chartered under the CIO Council.
Industry Speakers

Chid Apte, Ph.D.
Director, Analytics Research, Business Analytics & Mathematical Sciences
IBM

Dr. Chid Apte is Director of Analytics Research in the IBM Research Division, at the Thomas J. Watson Research Center in Yorktown Heights, New York. His organization is focused on Data Analytics and Business Analytics research. In addition to pursuing a core research agenda in Statistical Modeling & Forecasting, Machine Learning & Predictive Modeling, Data Mining Systems, Decision and Stochastic Analysis, and Massive Scale Analytics, his teams are engaged in several solutions initiatives focused on Workforce Analytics, Smarter Finance, Risk Management, Social Analytics, Spatio-Temporal Analytics, Customer and Market Intelligence, and Decision-Making under Uncertainty.

Dr. Apte has worked for over 25 years as a research scientist and technical leader in the data mining and business intelligence industry with significant experience in predictive analytics solutions. He has led several projects to develop analytics extensions for IBM’s Information Management and Business Analytics software product portfolio, and develop novel data mining solutions for IBM’s clients.

Dr. Apte received his Ph.D. in Computer Science from Rutgers University, and a B. Tech. in Electrical Engineering from the Indian Institute of Technology (Bombay). He is a member of the IBM Academy of Technology. He has published extensively in his areas of expertise, and actively involved in leading data mining conferences.

Robert Ames
Senior Vice President, Information and Communications Technologies
In-Q-Tel

Robert Ames is a Senior Vice President, Information and Communications Technologies at In-Q-Tel. In this role, he leads the practice in in setting the technical strategy, objectives, and investment directions to support the mission needs of the Intelligence Community. Robert and his team work closely with government leaders to identify, adapt, and deliver high value innovations from start-up companies to support vital government missions. His background is in the areas of

Prior to his role at In-Q-Tel, Robert was Director and Deputy CTO for IBM Federal. His role at IBM was to provide insight on his customer’s greatest challenges to IBM’s research and development organizations, and to build partnerships between IBM and its Federal clients. Those partnerships stretched technical boundaries and solved major problems as the government sought to modernize its computing and analytical capabilities, from the Navy’s shipboard environment, to the government’s largest data centers. Robert is a regular speaker at industry conferences, and is internationally recognized for his expertise in his Big Data and Cloud Computing.

Information technology is Robert’s second career. Prior to embarking on this career, he was a professional opera singer, most recently singing with The Washington Opera.

**Chip Copper, Ph.D.**

Principal Engineer  
Brocade

Dr. Copper is a storage area network expert with Brocade and has more than 30 years of experience in program management and information systems with a broad technical background. He has a Ph.D. from the University of Pittsburgh in Distributed Computing.

**Dave Denson**

Big Data Solution Architect  
NetApp

Dave Denson is an Electrical Engineer by education and a System Engineer by trade. He is a graduate of the University of Florida’s College of Engineering and has over 30 years of industry experience in the system engineering, management consulting, and software development areas. His sensor fusion work spans over 20 years, and his storage systems work spans over 12 years.
He has worked for Texas Instruments, Amdahl, Fujitsu, LSI, and several startups. He now works for NetApp's US Public Sector business as an ISR Solution Architect. David is life-long equestrian, and lives in Clifton, Virginia with his wife, 2 children, and a menagerie of dogs, cats, and horses.

Bob Kimball
Chief Technology Officer
Ciena Government Solutions

Mr. Kimball is the CTO of Ciena Government Solutions responsible for identifying technology directions and product positioning for sales of Ciena optical networking products in the DOD, Civilian Government, and Research and Education market segments. Recent projects include technology refresh of existing Government networks with 100G transport, the impact of Open Flow in the Government space, enhanced network security using carrier Ethernet technology and the impact to Green initiatives from the evolution of optical transport and switching technologies.

Prior to joining Ciena, Mr. Kimball held positions at the Defense Information Systems Agency, MITRE, and AT&T Bell laboratories where he was a subject matter expert and engineering manager in transmission engineering for both submarine and terrestrial optical networks. He has a background in nonlinear optics and modeling complex optical systems.

Steve LeSueur
Contributing Editor
1105 Government Information Group

Steve LeSueur provides Washington area businesses with public relations and communications services. He helps companies develop media strategy, writes press announcements and provides outreach to media. He writes and edits white papers, case studies and reports, and helps business executives write articles for publication. Steve also consults on website redesign and writes Web content.

Steve also serves as a contributing editor for custom media at 1105 Government Information Group, and is a regular moderator of 1105 webcast programs.
Before starting LeSueur Communications in August 2005, Steve served more than five years as editor of Washington Technology, a business magazine published by 1105 Media Inc. As editor, Steve directed the magazine’s coverage of companies that provide IT services to federal, state and local governments. This included both the Washington Technology website, which publishes daily stories, blogs and other features, and Washington Technology magazine, which provides in-depth coverage of the major business opportunities, events, legislation, policies and emerging technologies in the government IT market.


He holds a master’s degree in history from George Mason University, Fairfax, VA, and an undergraduate degree in history from Brigham Young University, Provo, UT. He wrote the award-winning book, “The 1838 Mormon War in Missouri,” published by the University of Missouri Press, Columbia, MO, and has published numerous articles on Mormon history.

Gary Newgaard
Director of Federal Isilon Business Unit
EMC Isilon

Gary Newgaard is the Director of Federal at EMC Isilon. Prior to joining EMC Isilon, Gary held senior positions at ASRC Management, Paragon Systems, and Compaq/HP. Gary earned his Bachelor’s of Science in Business Administration with a major in Marketing from the University of Tennessee. In addition to his university degree, Gary went on to the Executive Education Program at Harvard Business School.

Mr. Newgaard sits as a board member on various industry councils and is an entrepreneurial, results-driven leader with a strong and solid history of driving IT/high-tech market share, building revenue and profits, and bringing customer satisfaction to new heights. Gary resides in Northern VA and works out of the McLean, VA EMC office.

Clark Richey
Technical Director, Public Sector
MarkLogic
Clark is the Technical Director, Public Sector for MarkLogic, managing all Public Sector pre-sales engineering. Clark has an over twelve years experience in software engineering and system architecture with an emphasis on delivering large scale systems to Government customers.

Clark is an active MarkLogic Community member, contributing to several open source projects. In addition to working at MarkLogic, Clark teaches computer science and Big Data at Loyola’s master’s program. Clark has a B.S. in Computer Science from U.M.U.C. and a Masters in Computer Science from American University.

**Patrick Sims**

Senior Director  
Global Center of Excellence for Analytics, SAP

MISSING BIO

**Presentations**

Source: Email, June 27, 2013

Thank you for attending the FCW Executive Briefing on Managing Big Data for Results. As our guest, we hope that you found the seminar to be a productive, engaging forum and we are looking forward to seeing you at future events!

**Click the links below for instant access to the speaker presentations:**

**Applying Analytics to Detect Fraud and Ensure Compliance**  
C. Scot Atkins, Program Manager, Supercomputing and Revenue Protection, U.S. Postal Service

**Data to Decision: What Your Data Can Do For Your Agency**  
Patrick Sims, Senior Director, Global Center of Excellence for Analytics, SAP

**Control, Manage, and Analyze Big Data—Practical Approaches**  
Dave Denson, Big Data Solution Architect, NetApp
Using Agency Big Data to Transform Operations
Gary Newgaard, Director of Federal Isilon Business Unit, EMC Isilon

Mastering the Art and Science of Big Data Management
Robert Kimball, Chief Technology Officer, Ciena Government Solutions

How Big Data Applications Support Streamlined Operations and Risk Management
Clark Richey, Technical Director, Public Sector, MarkLogic

Please visit the links below to learn more about Big Data courtesy of our sponsors:

Learn more about the Brocade Federal Forum.

Learn more about the EMC Isilon Scale-out Storage Platform.

Analyst’s Notebook Premium Datasheet by IBM

Intelligence Analysis Platform by IBM

Creating an Actionable View of Data with Data Virtualization Whitepaper by MarkLogic

Learn more about MarkLogic and Download for free – MarkLogic Developer License

Solving Agencies’ Big Data Challenges: PED for On-the-Fly Decisions Whitepaper by NetApp

Reframing Big Data Whitepaper by SAP