# Operational Choices in Generating Real Time Political Event Data

Philip A. Schrodt, Ph.D.

Parus Analytics LLC and Open Event Data Alliance
Charlottesville, Virginia USA
http://philipschrodt.org
https://github.com/openeventdata/

Institute for Research on Statistics and its Applications
and Department of Political Science
University of Minnesota
24 September 2018

PARUS
ANALYTICS

# Event Data: Core Innovation

Once calibrated, monitoring and forecasting models based on real-time event data can be run [almost...] entirely without human intervention

- Web-based news feeds provide a rich multi-source flow of political information in real time
- Statistical and machine-learning models can be run and tested automatically, and are 100% transparent

In other words, for the first time in human history we can develop and validate systems which provide real-time measures of political activity without any human intermediaries

# Major phases of event data

- 1960s-70s: Original development by Charles McClelland (WEIS; DARPA funding) and Edward Azar (COPDAB; CIA funding?). Focus, then as now, is crisis forecasting.
- 1980s: Various human coding efforts, including Richard Beale's at the U.S. National Security Council, unsuccessfully attempt to get near-real-time coverage from major newspapers
- 1990s: KEDS (Kansas) automated coder; PANDA project (Harvard) extends ontologies to sub-state actions; shift to wire service data
- early 2000s: TABARI and VRA second-generation automated coders; CAMEO ontology developed
- 2007-2011: DARPA ICEWS project
- 2012-present: full-parsing coders from web-based news sources: open source PETRARCH coders and proprietary Raytheon-BBN ACCENT coder

# News Story Example: Example: 18 December 2007

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

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

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

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

# TABARI Coding: Lead sentence

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

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: First event

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

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Actors

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

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Agent

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

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Second event

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

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Second event target

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

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# TABARI Coding: Agent

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

Event Code: 111
Source: IRQ GOV
Target: TUR

Event Code: 223
Source: TUR
Target: IRQKRD REB

# Development of event ontologies

1970s: WEIS, COPDAB, CREON and others

1980s: BCOW (Leng) (crisis data: 300 categories)

1990s: PANDA (Bond): first ontology to focus on substate actors

2000s: IDEA (Bond, VRA): backward compatible with multiple existing ontologies, adds non-political events such as disaster and disease

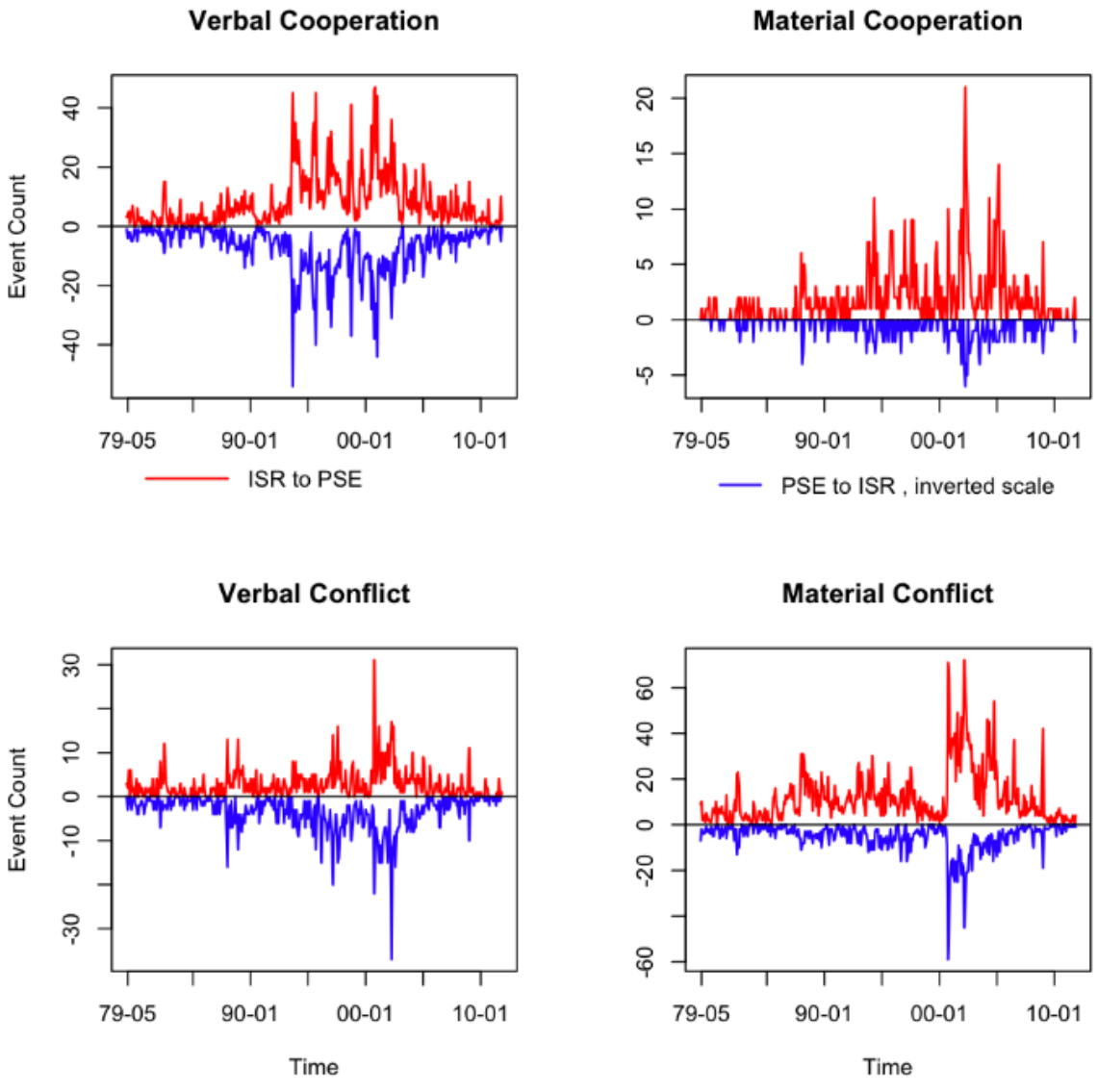
2000s: CAMEO (Gerner and Schrodt): combines ambiguous WEIS categories, expands violence and mediation-related categories; implemented as 15,000-phrase TABARI dictionary

late 2010s: PLOVER: generalized political coding scheme and data interchange specification

# WEIS primary categories (ca. 1965)

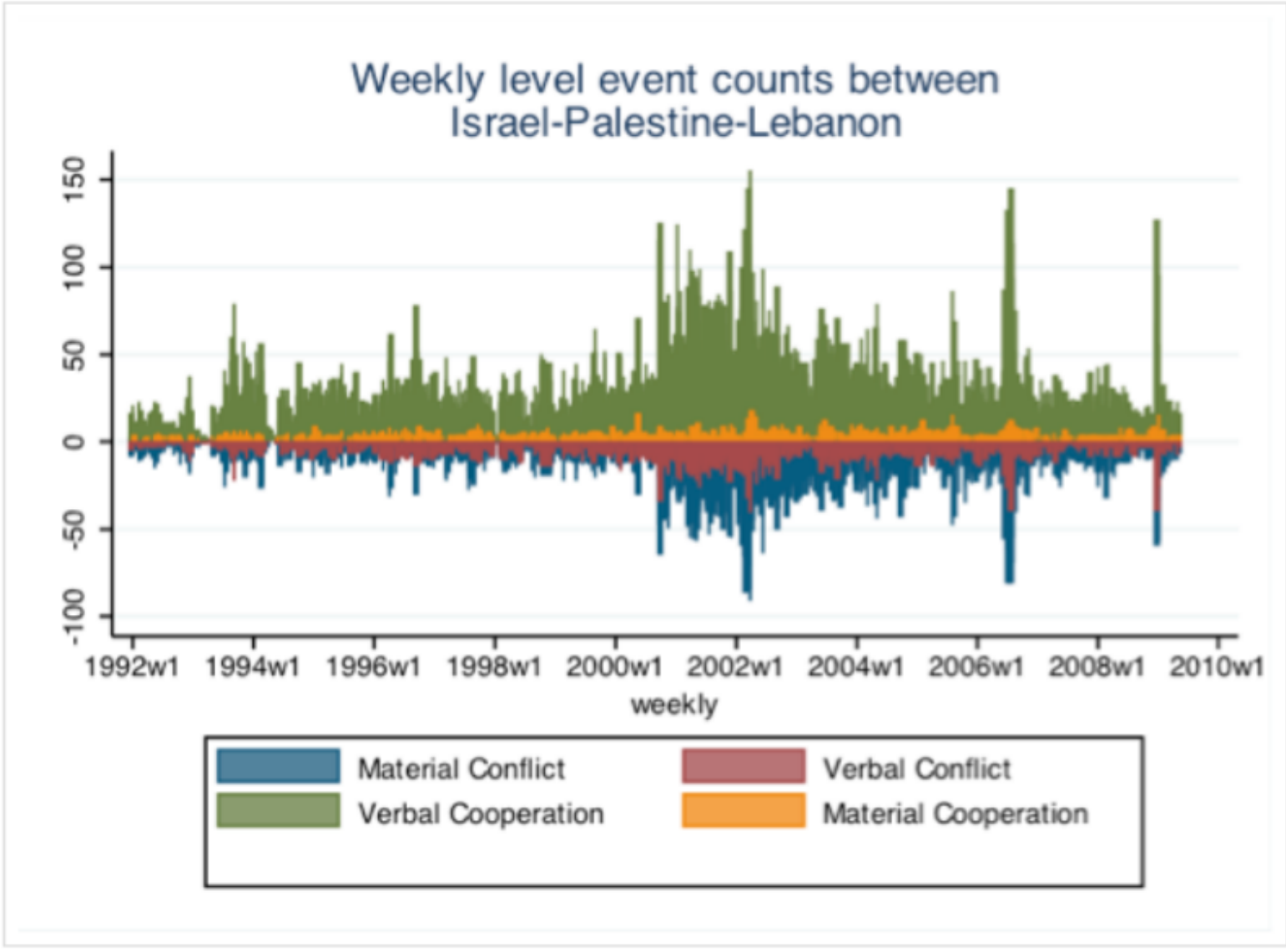
| | | | |
|---|---|---|---|
| 01 | Yield | 11 | Reject |
| 02 | Comment | 12 | Accuse |
| 03 | Consult | 13 | Protest |
| 04 | Approve | 14 | Deny |
| 05 | Promise | 15 | Demand |
| 06 | Grant | 16 | Warn |
| 07 | Reward | 17 | Threaten |
| 08 | Agree | 18 | Demonstrate |
| 09 | Request | 19 | Reduce Relationship |
| 10 | Propose | 20 | Expel |
| | | 21 | Seize |
| | | 22 | Force |

# KEDS Project Levant Data, 1979-2010

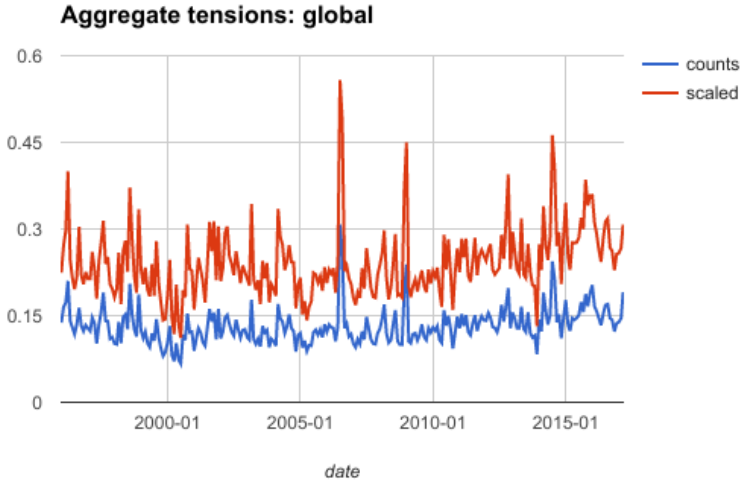


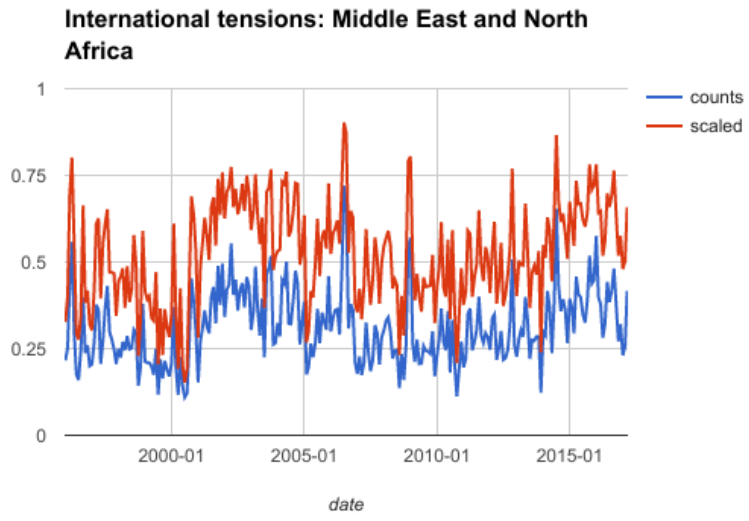
# KEDS Project Levant Data, 1992-2010

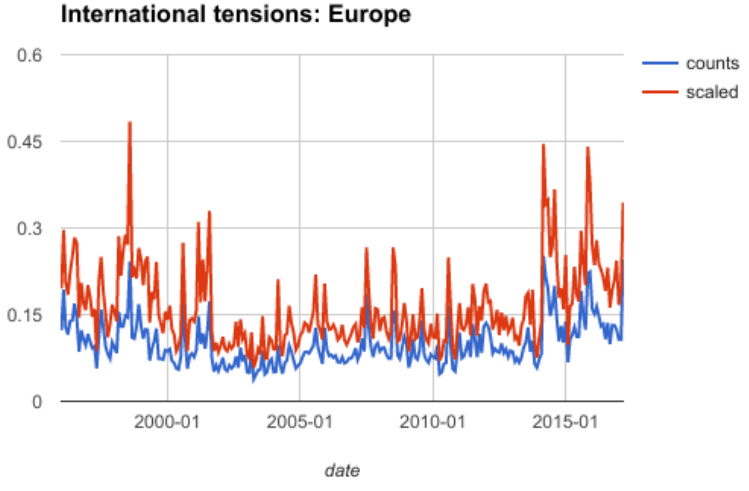
Visualization by Jay Yonamine (Penn State Political Science Ph.D. 2013, now Head of Data Science for Global Patents at Google)

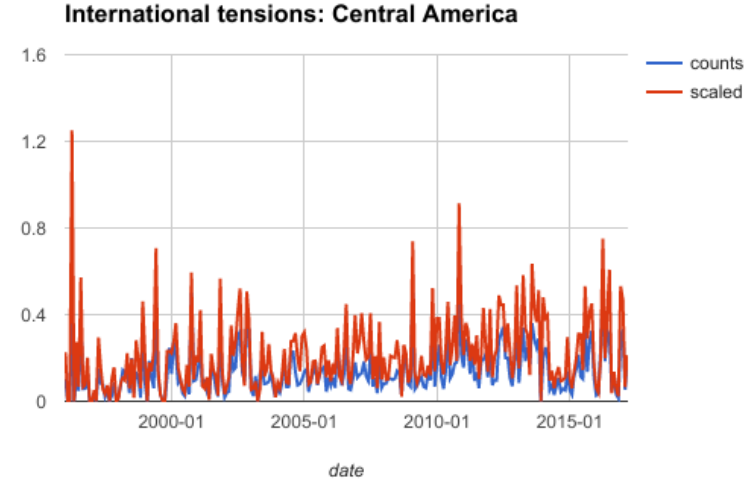


# Indicators derived from ICEWS, 1996-2017

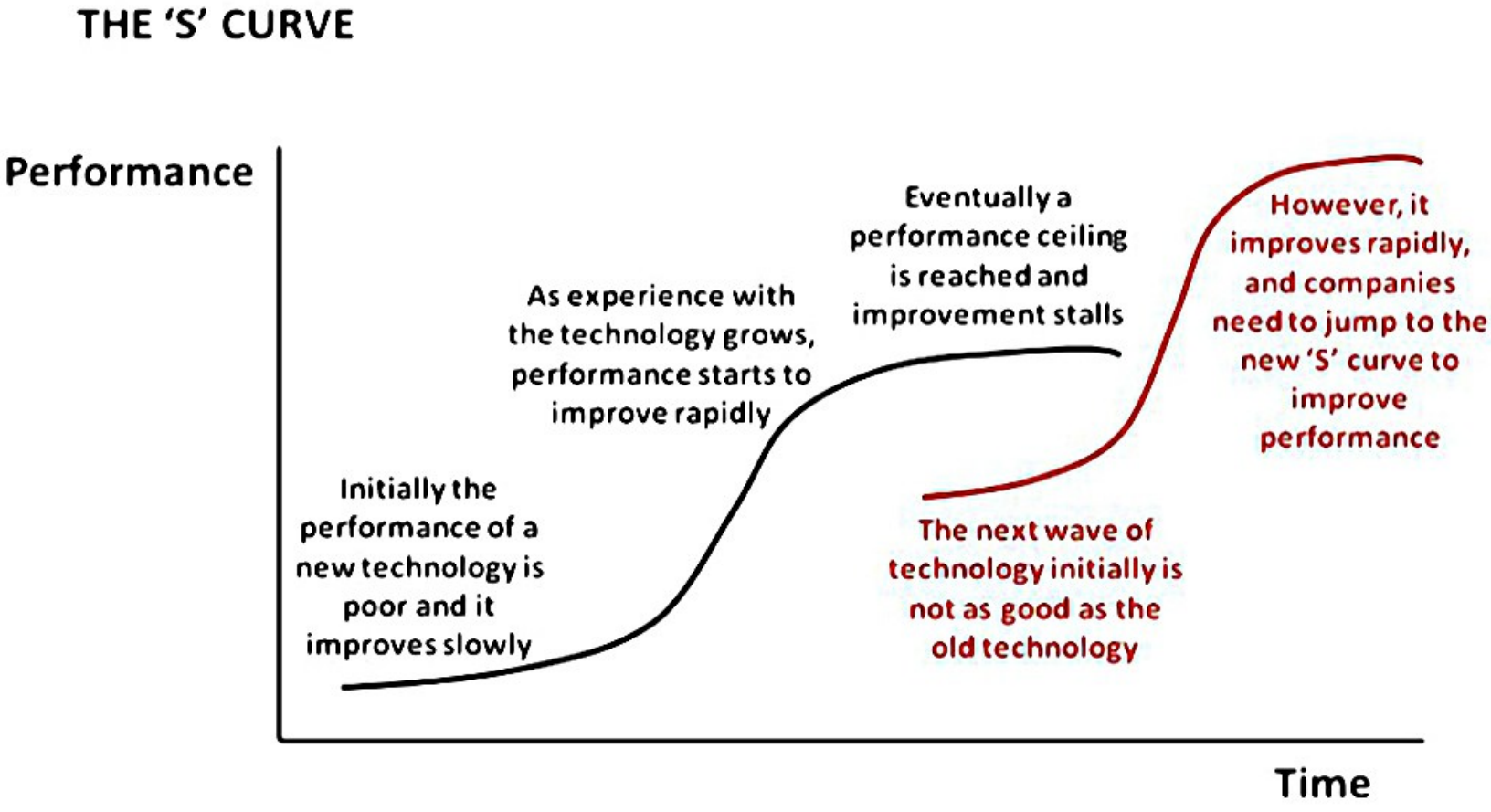


# Is event data ready for disruption?



# Are we at the flat point on a lower S-curve?

- David Honey (DARPA/ODNI) notes that hype is maximized when the curve flattens: please note that at present most people think event data sucks
- Machine coding did a classical disruption on human coding because it was lower quality but cheaper: in Clayton Christensen's theory this drives S-curve disruptions.
- Machine learning classifiers—support vector machines or neural networks—might replace patterns/dictionaries as cheaper-not-better if gold standard records (GSRs) become available. This has been done on toy problems.
- S-curves can level off and stay there:
  - Diesel locomotives
  - Boeing 737
  - 70-mph highway speed limit

# Another take on this

- IARPA PM at recent meeting: "I've talked to lots of analysts: no one has any use for event data."
- Twelve hours later, same meeting, a government analyst: "We *love* your event data tension model!" Suggesting the issue is open.
- Observation: Event data never really takes off—in either government or academic research—but it also never goes away: see `http://openeventdata.org/datasets.html` which lists 16 active projects.
- Observation: For the first time in the history of the field, the most innovative work has shifted to Europe—VIEWS, GCRI, ACLED, EMM

# Another take on this

- An IARPA PM at recent meeting: "I've talked to lots of analysts: no one has any use for event data."
- Twelve hours later, same meeting, a government analyst: "We *love* your event data tension model!" Suggesting the issue is open
- Observation: Event data never really takes off—in either government or academic research—but it also never goes away: see `http://openeventdata.org/datasets.html` which lists 16 active projects.
- Observation: For the first time in the history of the field, the most innovative work has shifted to Europe—VIEWS, GCRI, ACLED, EMM. These slides are based on talks I've given this year in Berlin and Brussels, not Washington.

# Overview of operational issues

Most of the infrastructure required for the automated production of political event data is now available through commercial sources and open-source software developed in other fields: it no longer needs to be developed specifically for event event production. However, a number of open questions remain:

- OEDA experience in the difficulties of maintaining a cloud-based software pipeline
- Maximizing vs "white-listing" news sources
- Coding ontology: weaknesses in CAMEO
- Approaches to multi-language coding
- Open source versus closed software solutions

# Challenges discovered in OEDA's "Phoenix" project

Real time data is easy to get *started*—we have multiple software pipelines available on GitHub—but *keeping it running* is a challenge...

- Cloud services are still evolving
- We selected an unreliable (but inexpensive!) provider which required periodic reboots: we eventually had to abandon this.
- Filtering, even for white-listed sources, needs to be robust
- We over-estimated the maturity of our coding program, PETRARCH-2, and didn't provide systematic dictionary updates
- As a volunteer organization, maintaining continuity when individuals moved to new responsibilities was difficult

Phoenix is currently hosted through a U.S. National Science Foundation project at the University of Texas/Dallas, but that funding ends in early 2019.

# Maximizing vs "white-listing" news sources

OEDA has deliberately chosen *not* to maximize the number of sources we code:

- Coding "everything" is surprisingly demanding in terms of computing resources, particularly when computationally-intensive parsing and/or translation is involved
- Obscure sources with unconventional editing are likely to cause coding errors and increase demands on dictionaries
- Censorship, rumors and "fake news" are a serious issues
- Most applications of event data rely on central tendencies, not finding a "needle in haystack"

Systematic research needs to be done on what, if anything, is gained from sources beyond those commonly used: the number of events generated by ICEWS drops off steeply beyond about twenty high-frequency "main-stream media" sources.

# Possible news sources

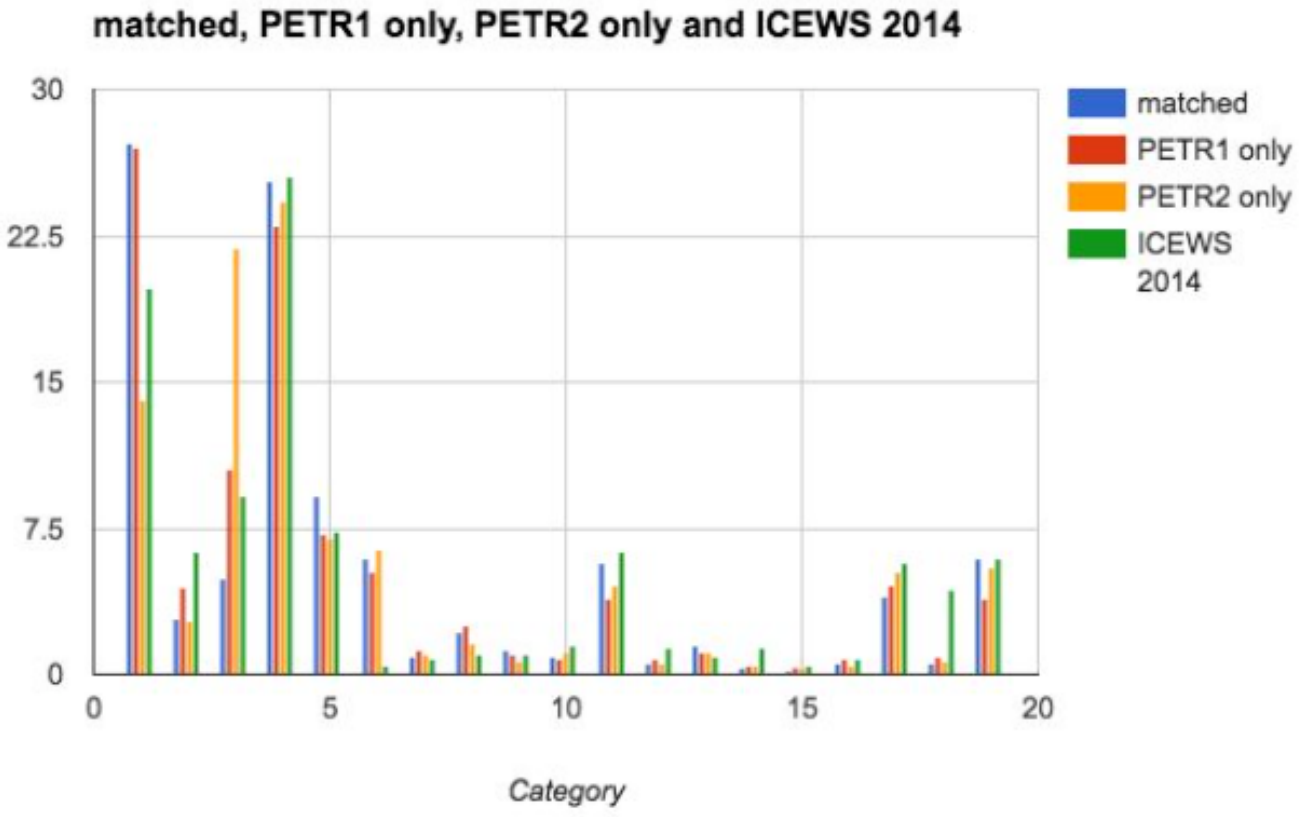
- International news services: most common sources for most data; quality is fairly uniform but attention varies
- Local media: quality varies widely depending on press independence, local elite control, state censorship, and intimidation of reporters
- Local NGO networks: these can provide very high quality information but require extended time and effort to set up
- Social media: These can be useful in very short term (probably around 6 to 18 hours) but have a number of issues
  - most content is social rather than political
  - bots of various sorts produce large amount of content
  - difficult to ascertain veracity: someone in Moscow or Ankara may be pretending to be in Aleppo

# Coding schemes: WEIS primary categories (ca. 1965)

| | | | |
|---|---|---|---|
| 01 | Yield | 11 | Reject |
| 02 | Comment | 12 | Accuse |
| 03 | Consult | 13 | Protest |
| 04 | Approve | 14 | Deny |
| 05 | Promise | 15 | Demand |
| 06 | Grant | 16 | Warn |
| 07 | Reward | 17 | Threaten |
| 08 | Agree | 18 | Demonstrate |
| 09 | Request | 19 | Reduce Relationship |
| 10 | Propose | 20 | Expel |
| | | 21 | Seize |
| | | 22 | Force |

This was updated around 2002 into the CAMEO system, which is used in all of the systems in the United States. However, CAMEO was explicitly designed for the study of international mediation, not as a general-purpose political event ontology.

# "CAMEO-World" across coders and news sources




Between-category variance is massively greater than the between-coder variance.

# PLOVER

Political Language Ontology for Verifiable Event Records

Event, Actor and Data Interchange Specification

Open Event Data Alliance

http://openeventdata.org/

http://ploverdata.org/

DRAFT Version: 0.6b2

March 2017

# PLOVER objectives

- Only the 2-digit event "cue categories" have been retained from CAMEO. These are defined in greater detail than they were in WEIS and CAMEO.

- Some additional consolidation of CAMEO codes, and a new category for criminal behavior

- Standard optional fields have been defined for some categories, and the "target" is optional in some categories.

- A set of standardized names ("fields") for line-delimited JSON (`http://www.json.org/`) records are specified for both the core event data fields and for extended information such as geolocation and extracted texts;

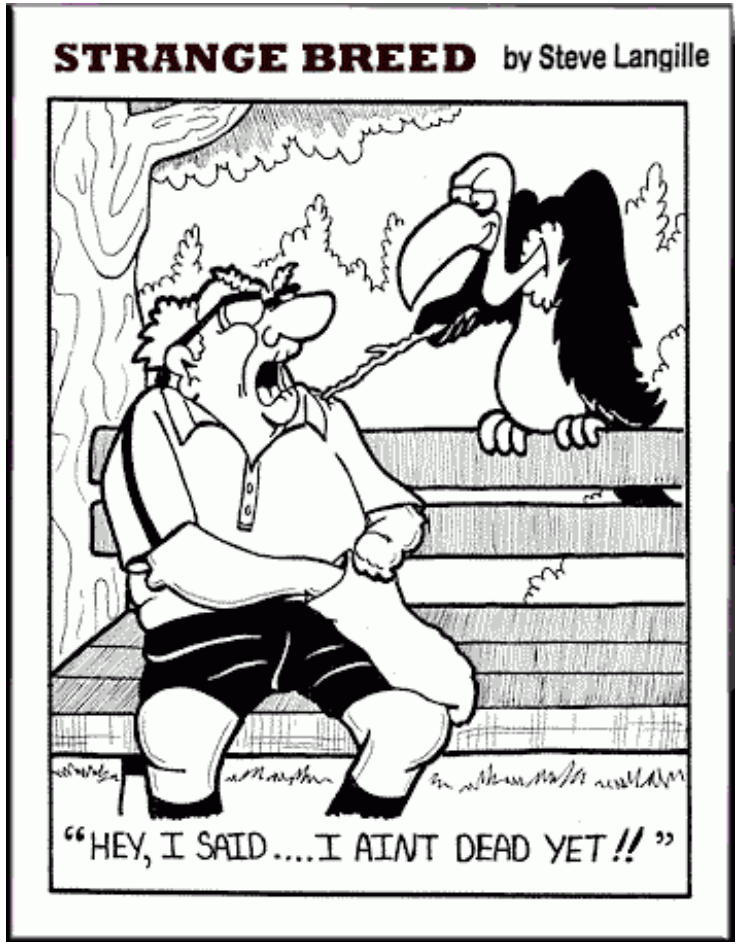
- We have converted all of the examples in the CAMEO manual to an initial set of English-language "gold standard records" for validation purposes—these files are at `https://github.com/openeventdata/PLOVER/blob/master/PLOVER_GSR_CAMEO.txt`—and we expect to both expand this corpus and extend it to at least Spanish and Arabic cases.

# Event, Mode, and Context

Most of the detail found in the 3- and 4-digit categories of CAMEO is now found in the *mode* and *context* fields in PLOVER. More generally, PLOVER takes the general purpose "events" of CAMEO (as well as the earlier WEIS, IDEA and COPDAB ontologies) and splits these into "$event - mode - context$" which generally corresponds to "$what - how - why$." We anticipate at least four advantages to this:

1. The "$what - how - why$" components are now distinct, whereas various CAMEO subcategories inconsistently used the *how* and *why* to distinguish between subcategories.
2. We are probably increasing the ability of automated classifiers—as distinct from parser/coders—to assign *mode* and *context* compared to their ability to assign subcategories.
3. In initial experiments, it appears this approach is *much* easier for humans to code than the hierarchical structure of CAMEO because a human coder can hold most of the relevant categories in working memory (well, that and a few tables easily displayed on a screen)
4. Because the words used in differentiate *mode* and *context* are generally very basic, translations of the coding protocols into languages other than English is likely to be easier than translating the subcategory descriptions found in CAMEO.

# Dictionary-based coding



# Dictionary-based coding: Hey, I'm ain't dead yet!

- Language model of the parser involves thousands of hours of experimentation across multiple major NLP research projects across decades
- PETRARCH-2 and Raytheon/BBN's ACCENT/Serif have an explicit language model for political events
- Models of language subcomponents such as dates, locations, and named entities
- Two decades of human-coded dictionary development from the KEDS and TABARI projects
- The WordNet synonym sets, again the product of thousands of hours of effort
- A variety of very large data sets such as `rulers.org`, CIA World Leaders and Wikipedia for named-entity resolution

# Approaches to multi-language coding

- Ignore it on the assumption that most relevant events will be available somewhere in English, e.g. on `/en/` branches of major news web sites. This could be tested: I suspect English is sufficient for many regions but not Latin America and possibly not for Arabic and Chinese.
- Native language dictionaries: UT/Dallas RIDIR project is producing these for Arabic and Spanish, and has developed tools for assisting on this. These are highly labor intensive.
- "Bag of words" machine-learning approaches such as support vector machines, neural networks, and word-embedding approaches (Google's `Word2Vec`). These require a large number of training cases.
- Machine translation: systematic experiments are needed here, and obviously the technology is rapidly improving

# Conjecture on multi-language coding

Machine translation (MT) in 2018 is where real-time mapping software was in early 2008, just after first iPhone : best systems were costly, though new free systems were workable

As with real-time mapping, MT is nearing (or past) the S-curve "take-off" point where the speed will improve dramatically while cost drops; quality has already improved substantially due to deep learning approaches. E.g. EMM recently developed a high-volume MT system for 17 languages into English optimized for news articles. There's more to MT than Google.

It is very, very difficult to envision a scenario where the resources available for the dictionary improvements in general-purpose native language event coders will produce results superior to improvements in MT, except possibly in some specialized applications.

# Open versus proprietary software

I'm not exactly a neutral observer on this issue...

- The open source environment for both natural language processing and event coding is now extraordinarily rich and largely has standardized on the Python programming language. It is thoroughly international.
- Open source software is nonetheless only "free as in puppy:" very substantial investment of labor is required to effectively use a complex open source system
- Continued maintenance and documentation of an open source system depends on the development of a large user community: there are serious network effects in operation
- There may still be some institutional resistance to open source

# Similar issues in. . . astrophysics

## Computational astrophysics for the future

**Simon Portegies Zwart**
+ See all authors and affiliations

*Science* 07 Sep 2018:
Vol. 361, Issue 6406, pp. 979-980
DOI: 10.1126/science.aau3206



Scientific discovery is mediated by ideas that, after being formulated in hypotheses, can be tested, validated, and quantified before they eventually lead to accepted concepts. Computer-mediated discovery in astrophysics is no exception, but antiquated code that is only intelligible to scientists who were involved in writing it is holding up scientific discovery in the field. A bold initiative is needed to modernize astrophysics code and make it transparent and useful beyond a small group of scientists.

# Similar issues in. . . astrophysics

## Computational astrophysics for the future

**Simon Portegies Zwart**

+ See all authors and affiliations

*Science* 07 Sep 2018:
Vol. 361, Issue 6406, pp. 979-980
DOI: 10.1126/science.aau3206



Scientific discovery is mediated by ideas that, after being formulated in hypotheses, can be tested, validated, and quantified before they eventually lead to accepted concepts. Computer-mediated discovery in astrophysics is no exception, but antiquated code that is only intelligible to scientists who were involved in writing it is holding up scientific discovery in the field. A bold initiative is needed to modernize astrophysics code and make it transparent and useful beyond a small group of scientists.


Obligatory picture of animal!

# Can't resist sharing this...

**dinosource**

Astrophysics phrase for poorly documented laboratory software written on the assumption it would only be used for a couple years but still in use, typically endlessly patched, and by multiple projects, two or three decades later.

# Issues for astrophysics software relevant to event data

- Open access to source code is essential for scientific progress and integrity: "secretly developed codes are of no help to the community and produce unverifiable results."
  - Not doing well here: Cline Center, TERRIER and Phoenix are coded with open PETRARCH-2 but the more widely used ICEWS and GDELT use secret coding engines
- Open standards for interchange of program parameters
  - Reasonably okay: ICEWS actor dictionaries are open if odd; TABARI/PETRARCH family is a de facto standard
- Modularized—LEGO blocks—components
  - Doing very well here with modular formatters, parsers, coders, geolocation, pipelines
- Core components need to be available that have been written and documented to industry standards, not laboratory standards
  - Still needs work: PETRARCH family has very poor documentation; TABARI/JABARI and Serif/ACCENT had professional programming, though only TABARI is open

# Open Event Data Alliance software

**Birdcage**

Basic, Integrated, and Reliably Distributed Coding, Actors, and Geolocation for Events

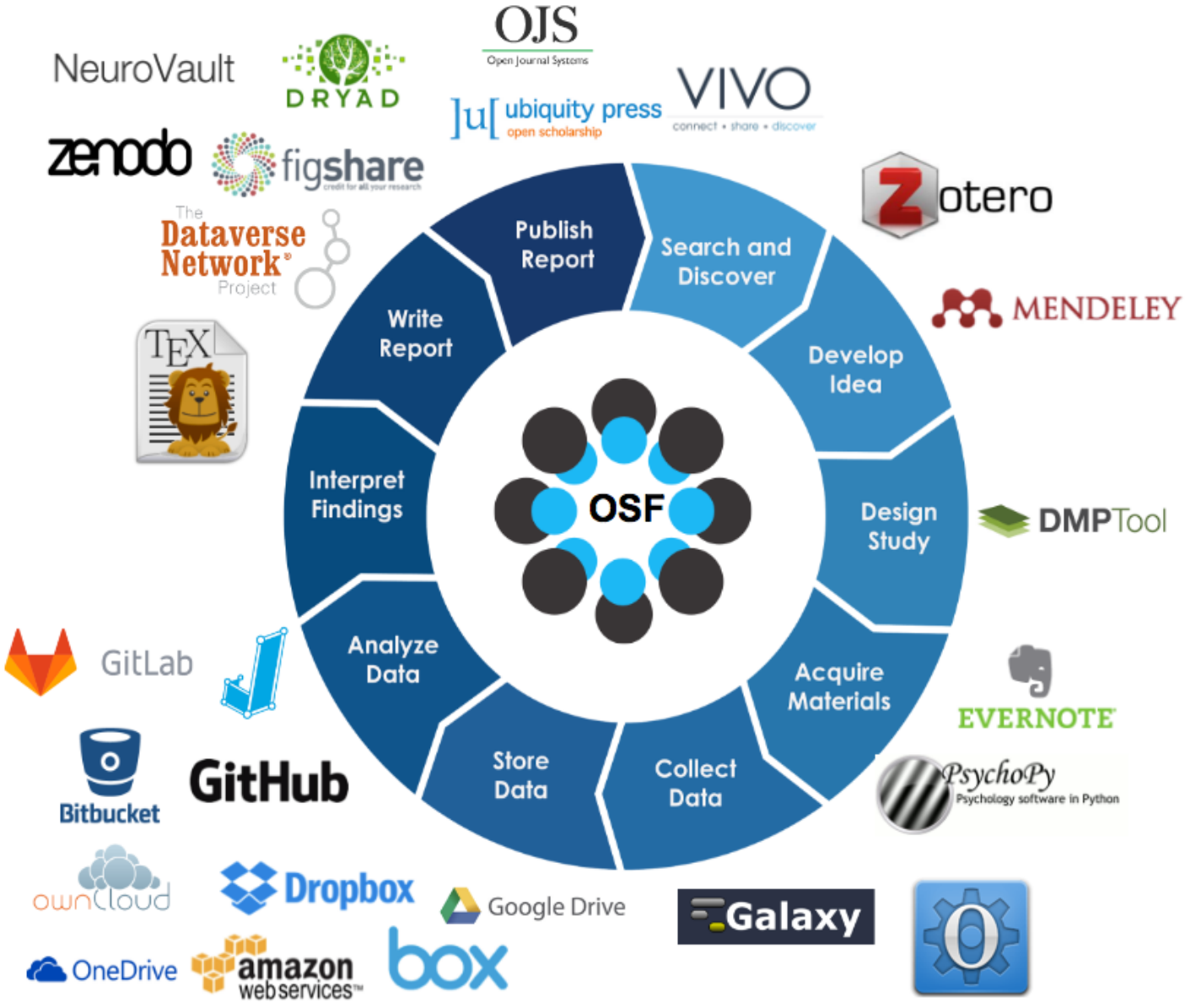
PETRARCH family of automated event data coders and dictionaries for CAMEO ontology

PLOVER Event Data Ontology


FJOLTYNG: PLOVER- and universal dependency-based event coder


PETRARCH-based web scraping and event coding pipeline

# Probably need to go beyond just GitHub...



# Remaining challenges: gold standard records

These are essential for developing example-based machine-learning systems but are extremely expensive to produce using existing methods

- They would allow the relative strengths of different coding systems to be assessed: FWIW this turns out to be essential for academic computer science publications
- We don't want "one coder to rule them all": different coders and dictionaries will have different strengths because the source materials are very heterogeneous.

Alternatives

- "Bronze standard" records using high through-put machine-assisted binary annotators such as `prodigy`
- Automatic extraction of patterns from the hundreds of thousands of existing CAMEO-coded records

# Remaining challenges: source texts

It would be very useful to have an open text corpus similar to GigaWord covering perhaps 2000 to the present. This is useful for

- Robustness checks of new coding systems
- Tracking actors who were initially obscure but later become important
- Tracking new politically-relevant behaviors such as cyber-crime and election hacking

# Remaining challenges: institutional

- Absence of a "killer app": we have yet to see a "I absolutely must have one of those!" moment.
  - Commercial applications such as Cytora (UK) and Kensho (USA) are still low-key and below-the-radar.
- Sustained funding for professional staff
  - (IMHO) Academic incentive structures are an extremely inefficient and unreliable method for generating well-documented, production-quality software.
  - Community is too small and specialized for crowd-sourced support on StackOverflow and GitHub
  - 24/7/365 real-time systems occasionally break for unpredictable reasons, and need to have expert supervision even though they mostly run unattended
  - Updating and quality-control on dictionaries is essential and is best done with long-term (though part-time) staff
  - This effort could easily be geographically decentralized

# Thank you

Email:
schrodt735@gmail.com

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

Links to open source software:
`https://github.com/openeventdata/`

Links to *lots* of event data sites:
`http://openeventdata.org/datasets.html`

# Supplementary Slides

# Event data coding programs

- TABARI: C/C++ using internal shallow parsing; 160-page manual.

  `http://eventdata.parusanalytics.com/software.dir/tabari.html`

- JABARI: Java extension of TABARI : alas, abandoned and lost following end of ICEWS research phase
- DARPA ICEWS: Raytheon/BBN ACCENT coder can now be licensed for academic research use
- Open Event Data Alliance: PETRARCH 1/2 coders, Moredcai geolocation. `https://github.com/openeventdata`
- NSF RIDIR Universal-PETRARCH: multi-language coder based on dependency parsing with dictionaries for English, Spanish and Arabic
- Numerous experiments in progress with classifier-based and full-text-based systems

# PLOVER output


```
{
    "id": "test-0056-0036_1",
    "date": "2015-02-12",
    "source": [{"actorText": "Russian Foreign Minister Sergei Lavrov", "code": "RUS", "sector": "GOV"},
               {"actorText": "Iranian counterpart Mohammad Javad Zarif", "code": "IRN"}],
    "target": [{"actorText": "Syria crisis", "code": "SYR"}],
    "event": "DISCUSS",
    "eventText": "discussed",
    "mode": "mode-holder",
    "context": "context-holder",
    "text": "MOSCOW: Russian Foreign Minister Sergei Lavrov and his Iranian counterpart Mohammad Javad
     Zarif discussed the Syria crisis by phone Wednesday, the Russian Foreign Ministry said in a statement
    "language": "en",
    "publication": "mudflat test data",
    "coder": "Parus Analytics",
    "version": "0.5b1",
    "dateCoded": "2017-03-20",
    "comment": "test output from mudflat",
},
```

# PLOVER: ASSAULT modes

| Name | Content |
|---|---|
| beat | physically assault |
| torture | torture |
| execute | judicially-sanctioned execution |
| sexual | sexual violence |
| assassinate | targeted assassinations with any weapon |
| primitive | primitive weapons: fire, edged weapons, rocks, farm implements |
| firearms | rifles, pistols, light machine guns |
| explosives | any explosive not incorporated in a heavy weapon: mines, IEDS, car b |
| suicide-attack | individual and vehicular suicide attacks |
| heavy-weapons | crew-served weapons |
| other | other modes |

Adapted from Political Instability Task Force Atrocities Database:
`http://eventdata.parusanalytics.com/data.dir/atrocities.html`

# PLOVER: general contexts

| Name | Content |
|---|---|
| political | political contexts not covered by any of the more specific categories below |
| military | military, including military assistance |
| economic | trade, finance and economic development |
| diplomatic | diplomacy |
| resource | territory and natural resources |
| culture | cultural and educational exchange |
| disease | disease outbreaks and epidemics |
| disaster | natural disaster |
| refugee | refugees and forced migration |
| legal | national and international law, including human rights |
| terrorism | terrorism |
| government | governmental issues other than elections and legislative |
| election | elections and campaigns |
| legislative | legislative debate, parliamentary coalition formation |
| cbrn | chemical, biological, radiation, and nuclear attacks |
| cyber | cyber attacks and crime |
| historical | event is historical |
| hypothetical | event is hypothetical |

# Simple models are good!

Recent study on predicting criminal recidivism showed equivalent results could be obtained from

- A proprietary 137-variable black-box system costing $22,000 a year
- Humans recruited from Mechanical Turk and provided with 7 variables
- A two-variable statistical regression model

For this problem, there is a widely-recognized "speed limit" on predictive accuracy of around 70% and, as with conflict forecasting, multiple methods can achieve this.

Source: *Science* 359:6373 19 Jan 2018, pg. 263; the original research is reported in *Science Advances* 10.1126/sciadv.aao5580 (2018)