OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

TEST AND EVALUATION IN ACE AND OSI

IARPA

LEADING INTELLIGENCE INTEGRATION

Jason Matheny
January 16, 2013

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

AGGREGATIVE CONTINGENT ESTIMATION (ACE)

LEADING INTELLIGENCE INTEGRATION

Program Goals:

- Significantly increase the accuracy, precision, and timeliness of analytic judgments about global events.
- Develop and test methods that generate estimates by eliciting, aggregating, and communicating the judgments of many widely-dispersed analysts
- Identify weights that maximize the accuracy of the aggregate estimate, based on individual, group, and problem attributes. E.g.:
  - Individual: past performance, education, experience, languages, travel, supporting arguments
  - Group: size, diversity
  - Problem: domain, rarity

Examples of forecasting questions:

- Will Iran consent to nuclear program talks before 1 January 2013?
- Will the Taliban begin official in-person negotiations with either the US or Afghan government by 1 August 2012?
- Will the UN Security Council pass a new resolution concerning Syria by 1 August 2012?
- Will Yousaf Raza Gillani vacate the office of Prime Minister of Pakistan before 1 October 2012?
- Will Serbia be officially granted EU candidacy before 1 September 2012?
- Who will be inaugurated as President of Russia in 2012?
- Will North Korea successfully detonate a nuclear weapon by 1 January 2013?
- By 1 October 2012, will Egypt officially announce its withdrawal from its 1979 peace treaty with Israel?
- Will Kim Jong-un attend an official meeting with any G8 head of government before 1 October 2012?

Performer Tasks

- Each performer is required to perform the same tasks:
  - Conduct research towards accurate forecasting of events
    - Elicitation of judgments
    - Aggregation of judgments
    - Communication of forecasts
  - Deliver forecasts for real-world events
  - Deliver prototypes of systems that generate and communicate those forecasts
- Performers differ in their technical approach to research and their strategy for prototype development.
- ACE test and evaluation takes place continuously as real-world events are observed.

T&E Tasks

1. Collect forecasting questions nominated by teams, IARPA, MITRE, and IC
2. Edit questions so they are precise and resolvable
3. Assign questions to all teams
4. Assign questions to MITRE-managed control groups
5. Receive daily forecasts from all teams for all assigned questions
6. Score forecasts for accuracy as events occur
   - Mean of daily quadratic (Brier) scores: $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{j=1}^{T_i}\left(o_i - p_{ij}\right)^2$
   - We compare to other proper scoring rules: logarithmic and spherical

Metrics & Milestones

Year:	1	2	3	4
Study Population	Recruited by performer			
Forecasting Problems	Unclassified, unconditional, chosen by IARPA	Unclassified, unconditional & conditional, chosen by IARPA		
Metrics	Brier (MQS)			
Year-end Milestone (% improvement over Government's ULinOP)	20	35	50	50+

IFP Resolution Language

- Will a foreign or multinational military force invade, enter or significantly* fire on Iran before 21 January 2013?
- Only an unwelcomed incursion of non-Iranian troops into Iranian soil qualifies as "invaded" or "entered." An invasion or entry of an Iranian embassy abroad does not constitute an invasion of or entry into Iran. Similarly, an incursion into Iranian territorial waters or airspace does not sufficiently constitute an invasion of or entry into Iran. A "foreign or multinational military force" refers to some recognized subset of a nation's (or multinational coalition's) official military. This definition of "military force" excludes quasi-military or paramilitary groups, such as insurgents, mercenaries, guerillas, "rebels," independent militias, or terrorist actors. Also outside the scope of "foreign or multinational military force" are alleged actions attributed to covert intelligence services or representatives thereof, unless those actions are expressly acknowledged by a sponsoring nation's government. *"Significantly fired on", however, is defined more broadly, to include bombs, missiles, chemical or other unconventional weapons or small arms fired on non-captive Iranian troops, soil, or naval vessels (including citizens within Iranian territory), military installations, or military vehicles (e.g., ships, subs, tanks, jets), such that at least 10 Iranians are killed. Iranian embassies abroad do not constitute Iran. "Before" should be interpreted to mean at some point prior to the end (23:59:59 ET) of the previous day. Outcome will be resolved based on reporting from BBC News or Reuters or Economist Online (http://www.bbc.co.uk/news/ or http://www.reuters.com/ or http://www.economist.com). If nothing is reported in these sources, then the "status quo" outcome typically will be assumed (i.e., no new hostile acts). Administrator reserves the right to use other sources as needed (e.g., CIA World Factbook, Wikipedia), provided those sources do not directly contradict concurrent event reporting from BBC News, Reuters, or Economist Online. In cases of substantial controversy or uncertainty, administrator may refer the question to outside subject matter experts, or we may deem the question invalid/ void.

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

OPEN SOURCE INDICATORS (OSI)

LEADING INTELLIGENCE INTEGRATION

Program Goals

- Develop and demonstrate technology to “beat the news” by providing early warning of significant societal events through continuous, real-time global monitoring of diverse, publicly available data
- Events of interest: civil unrest, mass gatherings, mass violence, political elections, disease outbreaks, economic crises.
- Detect changes in population-level behaviors through signals extracted from pervasive, publicly available social sensors: blogs, twitter, internet traffic, web search queries, traffic webcams, Wikipedia edits, financial markets, others

Performer Tasks

- Develop a system that automatically sends warnings (structured emails) to T&E team
  - Civil unrest
    - 6 categories of unrest – employment and wages, housing, energy and resources, other economic policies, other government policies, other
    - Violent vs. non-violent unrest
  - Elections and referenda
    - National, state, and municipal elections
  - Infectious human illness events
    - Rare diseases – Machupo, Cholera, Hantavirus, Yellow Fever
    - Influenza-like-Illness (ILI)
    - Pandemic influenza
  - Economic events
    - Significant increases or decreases in stock indices and currency exchanges
- Warning = {[Population, Event code, mm/dd/yy, (Country, State, City)], probability}
- Generate an audit trail that links warnings to data and helps analyst understand why the warnings were generated

T&E Tasks

- Create the Gold Standard Report (GSR): identify and encode all relevant events that occurred between Jan 2011 to Apr 2012
- From May 2012, update the GSR each month: identify and encode all new events found in searches
- Ingest performers' warnings into database
- Match warnings to events in the GSR
- Score each warning and calculate aggregated metrics, monthly
- Settle performer disputes

Event Types

01 – Civil Unrest	011 – Employment & Wages	0111 – Non-violent Civil Unrest
		0112 – Violent Civil Unrest
	012 – Housing	0121 – Non-violent Civil Unrest
		0122 – Violent Civil Unrest
	013 – Energy & Resources	0131 – Non-violent Civil Unrest
		0132 – Violent Civil Unrest
	014 – Other Economic Policies	0141 – Non-violent Civil Unrest
		0142 – Violent Civil Unrest
	015 – Other Government Policies	0151 – Non-violent Civil Unrest
		0152 – Violent Civil Unrest
	016 – Other	0161 – Non-violent Civil Unrest
		0162 – Violent Civil Unrest
	017 - Unspecified	0171 – Non-violent Civil Unrest
		0172 – Violent Civil Unrest
02 – Vote	021 – Election	0211 – President/Prime Minister
		0212 – Governor
		0213 – Mayor
	022 – Referendum	0221 – “Yes”
		0222 – “No”
03 – Infectious Human Illness	031 – Rare Diseases	0311 – Bolivian Hemorrhagic Fever (Machupo)
		0312 – Cholera
		0313 – Hantavirus
		0314 – Yellow Fever
	032 – Pandemic	
	033 – Influenza Like Illness (ILI)	
04 – Economy	041 – Stock Index	0411 – Stock Index Increases
		0412 – Stock Index Decreases
	042 – Currency Exchange	0421 – Currency Exchange Increases
		0422 – Currency Exchange Decreases

GSR Production

- OSI typologies
  - Adapted from IDEA and Tabari typologies used in political science
  - Extended to handle OSI events: voting, human disease, economic events
  - World Gazetteer provides standardized location encoding; Tabari defines the actor classes
- Event-specific search and encoding
  - Civil Unrest: typology-directed searches against Gold Standard Sources (GSS) - 63 news sources, primarily Spanish and Portuguese
  - Voting: unrestricted web searches and monitoring official sources
  - Rare Diseases: unrestricted web searches, official sources, ProMED
  - Influenza-Like Illness (ILI): case count time series from PAHO
  - Economic Data: Bloomberg stock indices and FOREX data - z-score to detect “spikes” in stock or foreign exchange data

GSR Statistics

Views	Count
Total Events	3007
Active Events	2912
Inactive Events	95
Active Event Updates	442
2011 Average Events/Month	89
2012 Average Events/Month	204
New Events in September 2012	483
Most Active Country	Mexico 409

Total events encoded

Inactive (retired) events reflect typology changes

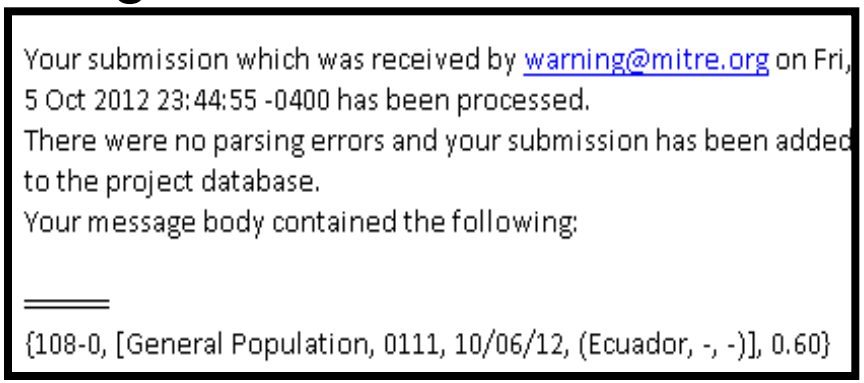
Updates due to source data revisions (e.g., PAHO updates), Performer feedback

Significant change in event/rate from 2011 to 2012. Potential factors: change to GSS, changes in the typology, country activity level, …

September rate > August > July >...

Warning Submission

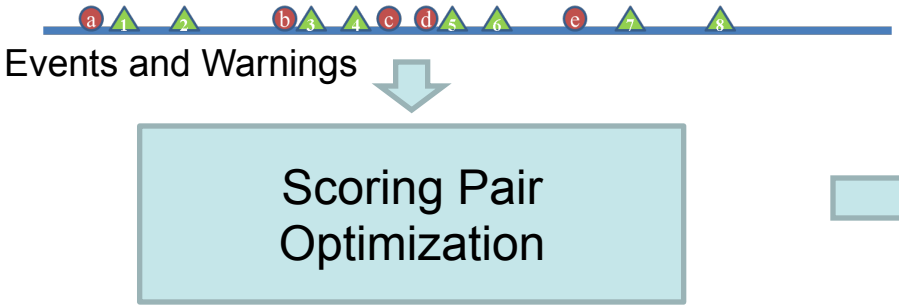
- Performers generate warnings
  - *{Warning ID, [Pop, OSI Event Code, mm/dd/yy, (Country, State, City)], prob}*
  - {108-0, [General Population, 0111, 10/06/12, (Ecuador, -, -)], 0.60}
- Sent via e-mail to warning@mitre.org with a common subject line: "SUBMISSION"
- Parsed and ingested into the Warning Database
- Confirmation message or error reports e-mailed to Performers
- Database stores the extracted warning, submission metadata (e.g., warning ID and alpha code), and date/time it was sent, received, and processed

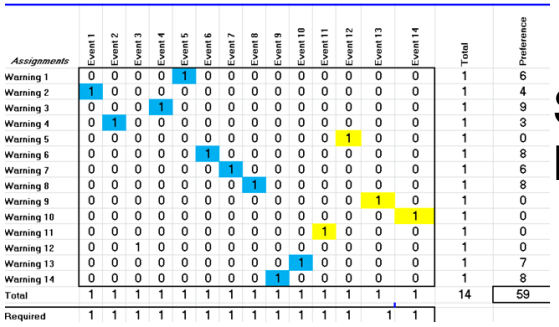

Your submission which was received by warning@mitre.org on Fri, 5 Oct 2012 23:44:55 -0400 has been processed.
There were no parsing errors and your submission has been added to the project database.
Your message body contained the following:

{108-0, [General Population, 0111, 10/06/12, (Ecuador, -, -)], 0.60}

Scoring Workflow

- Warnings are retrieved from the Warning Database and matched to candidate GSR records
- Best-match (Warning-GSR) algorithm addresses the many-to-many match problem and identifies the "best" warning-event pairs used in scoring:
  - Constrains matches so that each warning is matched to at most 1 event and each event is matched to at most 1 warning
  - Can be adapted for use with each event type
  - Addresses special cases (e.g., when there are more/fewer warnings than events)





Assignments	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6	Event 7	Event 8	Event 9	Event 10	Event 11	Event 12	Event 13	Event 14	Total	Preference
Warning 1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	6
Warning 2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4
Warning 3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	9
Warning 4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	3
Warning 5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
Warning 6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	8
Warning 7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	6
Warning 8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	8
Warning 9	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
Warning 10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
Warning 11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
Warning 12	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
Warning 13	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	7
Warning 14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	8
Total	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14	59
Required	1	1	1	1	1	1	1	1	1	1	1	1	1	1		

Scoring Pairs

- Warnings are matched to observed events and scored on three metrics
  - Lead time – # days between time the warning was sent and its earliest mention in the news
  - Probability score (Brier)
  - Warning quality – degree of match between the event description in the warning vs the GSR
- Aggregated metrics
  - Mean lead time, mean probability score, and mean quality score
  - Recall – (# of events in GSR for which team sent a warning)/ (total # of events in GSR)
  - Precision – (# of events in GSR for which team sent a warning)/ (total # of events the team submitted warnings for)

- **Lead Time**
  - **Days between warning and Gold Standard Report**
  - Warning for an event outside the 30-day window will not be scored, but should be submitted. Such events will be analyzed separately to provide additional assessment of the team's approach.
  - While successive, better warnings for the same event will be scored separately, teams will be asked to identify such successive warnings. The Government team will use this information for additional assessments of team's approach.

- **Probability Score**
  - **Quadratic score = $1 - (o-p)^2$**
  - p is the probability assigned to the warning, o is ground truth: 1 if the event occurred, 0 if the event didn't occur within 30 days.

- **For each warning we calculate the quality $q = \alpha_1+\alpha_2+\alpha_3+\alpha_4$**
  - α_1 ~ Population; α_2 ~ Event type; α_3~ Location; α_4 ~ Event time
  - This provides "partial credit" for partial warnings.
- **Use typology of populations, events, locations to calculate match between warning and ground truth**; e.g., for locations:
  - Typology = (Country, Province/State, City).
  - Compare warning location with true location to get (x_1, x_2, x_3), $x_i = 0$ if false, $x_i = 1$ if true.
  - Location quality = $\alpha_3 = 1/3\ x_1 + 1/3\ x_1x_2 + 1/3\ x_1x_2x_3$
    - If the warning has the wrong country, then $\alpha_3 = 0$.
    - If the warning has the country right but everything else wrong, $\alpha_3 = 1/3$.
    - If the warning has the country and the province/state right but the city wrong, $\alpha_3 = 2/3$.
    - If all is right, $\alpha_3 = 1$.
- **For the time of the event, use *1- min(|warning time – actual time|,30)/30***

Milestones

Metric	Month 12 (3 months of warnings)	Month 24 (12 months of warnings)	Month 36 (12 months of warnings)
Mean Lead Time	1 day	3 days	7 days
Mean Quality Score	3	3.25	3.5
Mean Probability Score	0.60	0.70	0.85
Recall	0.50	0.65	0.80
Precision	0.50	0.65	0.80