National Consortium for the Study of Terrorism and Responses to Terrorism
A Center of Excellence of the U.S. Department of Homeland Security

Event Data and the Construction of Reality

Brian Wingenroth, Erin Miller, Michael Jensen, Omi Hodwitz, Kieran Quinlan

Research Questions

Do automated, manual, and hybrid data collection techniques produce similar results?

If results differ, are those differences driven by systematic errors?

If these differences are driven by systematic errors, can we identify ways to address them?

Datasets

Global Database of Events, Language, and Tone (GDELT)

Global Terrorism Database (GTD-START)

Integrated Conflict Early Warning System (ICEWS)

Phoenix

Suicide Attack Database (SAD-CPOST)

Scope

Successful suicide bombings classified as terrorism that occurred worldwide in January and February 2015

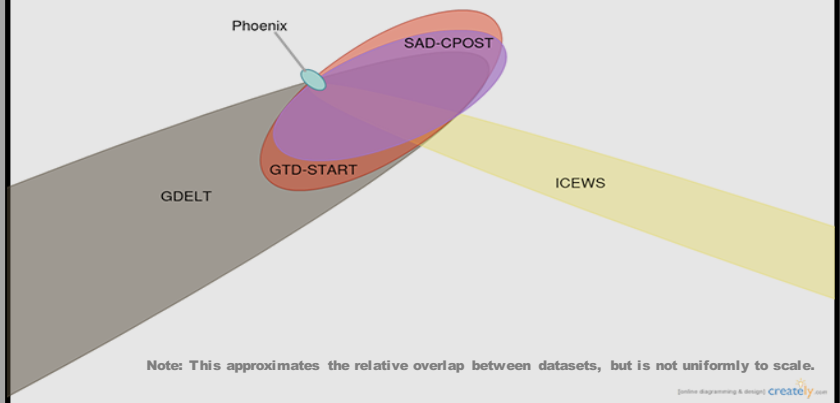
Methods

New dataset systematically matches all cases across five datasets, notes inconsistencies and reasons for inconsistencies

Metrics:

Precision; Recall; F-measures

Visualizing the Overlap



Note: This approximates the relative overlap between datasets, but is not uniformly to scale.

Metrics

	Suicide Events	% Duplicate	Precision	Recall	F-measure	F_2-measure
GDELT	5194	79%	0.02	0.70	0.04	0.08
GTD-START	116	0%	1.00	0.97	0.99	0.98
ICEWS	443	83%	0.17	0.57	0.27	0.39
Phoenix	10	65%	0.50	0.10	0.17	0.12
SAD-CPOST	103	5%	0.94	0.71	0.81	0.75

Definitions

- **"Ground truth": set union of unique suicide bombing events (n=136)**
- **Precision: percentage of suicide bombing events that are correct**
- **Recall: percentage of "ground truth" events recorded**
- **F-measure: harmonic mean of precision and recall**
- **F2 measure: modifies F-measure to emphasize recall**

Unique Contributions to "Ground Truth" (n=136)

GDELT ^ GTD ^ ICEWS ^ Phoenix ^ SAD	12	GDELT ^ GTD ^ ICEWS	10	GTD ^ Phoenix	1
GDELT ^ GTD ^ ICEWS ^ SAD	47	GTD ^ ICEWS ^ SAD	5	GTD	17
GDELT ^ GTD ^ ICEWS ^ Phoenix	1	GTD ^ SAD	14	ICEWS	2
GDELT ^ GTD ^ SAD	17	GDELT ^ GTD	8	SAD	2

Issues

Duplication of events:

- **Failure to remove identical articles**
- **Failure to detect updates to articles**
- **Failure to remove exact translations**

Inaccurate attributes:

- **Failed date resolution**
- **Failed location extraction**
- **Lack of topic segmentation**

False identification of events:

- **Poor source quality/validity**
- **Inclusion of irrelevant content**
- **Definitional inconsistencies**

Missing events (false negatives):

- **Limited source inclusion**
- **Lack of non-English sources**

Solutions

Pre-processing:

- **Filter articles for relevance**
- **Remove similar articles**
- **Model semantic overlap of articles**

Accuracy:

- **Improve date/time extraction**
- **Improve location extraction**
- **Accurately segment topics**

Verify and corroborate assertions

Conclusion

Data collection strategies dramatically impact resulting data. Human validation is critical and automated strategies should leverage existing and new technologies.

Acknowledgements

The Global Terrorism Database is funded through START, by the US Department of State (Contract Number: SAQMMA12M1292) and the US Department of Homeland Security Science and Technology Directorate's Office of University Programs (Award Number 2012-ST-061-CS0001, CSTAB 3.1). The views and conclusions contained in this presentation are those of the authors and should not be interpreted as necessarily representing the official views or policies of the United States Government.

References

National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2015). Global Terrorism Database [Data file]. Retrieved from http://www.start.umd.edu/gtd

Chicago Project on Security and Terrorism (CPOST). 2016. Suicide Attack Database (April 19, 2016 Release). [Data file]. Retrieved from http://cpostdata.uchicago.edu/