START

Innovative Algorithm and Database Development Relevant to Counterterrorism and Homeland Security Efforts at START

Report to the Office of University Programs,
Science and Technology Directorate,
U.S. Department of Homeland Security

August 2014

National Consortium for the Study of Terrorism and Responses to Terrorism
A Department of Homeland Security Science and Technology Center of Excellence
Based at the University of Maryland

8400 Baltimore Ave., Suite 250 •College Park,MD 20740 • 301.405.6600
www.start.umd.edu

About This Report

The authors of this report are Raul Garcia-Sanchez, Daniel Casimir and Prabhakar Misra, Department of Physics & Astronomy, Howard University, Washington, DC 20059. Questions about this report should be directed to Prabhakar Misra at pmisra@howard.edu.

START
NATIONAL CONSORTIUM FOR THE
STUDY OF TERRORISM AND RESPONSES TO TERRORIS

Raul Garcia-Sanchez, Prabhakar Misra, Daniel Casimir

This research was supported by the Department of Homeland Security Science and Technology Directorate's Office of University Programs through the Summer Research Team Program for Minority Serving Institutions. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or START.

About START

The National Consortium for the Study of Terrorism and Responses to Terrorism (START) is supported in part by the Science and Technology Directorate of the U.S. Department of Homeland Security through a Center of Excellence program based at the University of Maryland. START uses state-of-the-art theories, methods and data from the social and behavioral sciences to improve understanding of the origins, dynamics and social and psychological impacts of terrorism. For more information, contact START at infostart@start.umd.edu or visit www.start.umd.edu.

Contents

Abstract 3

Introduction 3

Goals 5

Methodology 6

Global Terrorism Database (GTD) 7

Contributions made to project 14

New Skills and Knowledge gained 14

The relevance of your research to the DHS mission 15

Profiles of Incidents involving CBRN by Non-state actors (POICN) database 16

Contributions made to the project 19

New skills and knowledge gained 20

Relevance of your project to the DHS mission 20

Presentations made to START Staff 20

Impact of Research Experience on Academic and/or Career Planning 21

Future Work 22

Classroom Plan 23

Acknowledgments 25

References 26

Abstract

The summer research project at START – National Consortium for the Study of Terrorism and Responses to Terrorism, a DHS Center of Excellence at the University of Maryland, College Park, June 2- August 8, 2014, involved pattern recognition in two large terrorism related databases, namely the Global Terrorism Database (GTD) and the Profiles of Incidents Involving CBRN by Non-State Actors (POICN) Database, employing neural networks and other machine learning algorithms. Training and test data subsets were extracted from previously compiled data to develop a variety of models based on machine learning pattern recognition algorithms that can enable prediction of future terrorist threats with specified percentage errors and uncertainties and thereby enable the intelligence and homeland security communities to make informed decisions regarding deployment of counter-terrorism resources in order to effectively thwart terror plots prior to their occurrence.

Introduction

The research project involved the development and utilization of methods that use machine learning to identify patterns in two terrorism-related databases, namely the Global Terrorism Database (GTD) and the Profiles of Incidents Involving CBRN by Non-State Actors (POICN) database (where CBRN is an acronym for Chemical, Biological, Radiological and Nuclear events). During the period 1990-present, there have been over 450 CBRN "events" ranging from plots to completed attacks using and involving a variety of sometimes rudimentary and often highly sophisticated technologies.

One of the primary objectives of the summer project was to find patterns from previously compiled data that can be utilized to predict behavior in newer data by developing a cluster-based algorithm. A central goal of the research project was to utilize neural networks and other machine learning algorithms to determine missing data from the Terrorism-related databases pioneered at the START DHS Center of Excellence (COE). For the current scope of the project, we focused on two such databases, the GTD and POICN. For example, data from CBRN attacks in the POICN database from 2000 could be used to teach the algorithm to generate output related to existing data in later years.

This, in turn, allowed us to discern patterns based on the multitude of variables used in the POICN data set by looking at links between (1) terrorist attacks based on region, (2) terrorist events and future attacks, and (3) different terrorist organizations. Additionally, although there are multitudes of terrorist organizations worldwide, which sometimes come and go, the methods developed here can be used to determine similar control structures and umbrella bodies that may be present among the various terrorist organizations, which may sometimes share individuals, and our approach could determine if these transient groups are part of a bigger group or network based on their attack modus operandi. More importantly, some data on past terrorist attacks might be directly connected to future terrorist acts through similar recognizable control structures; this might be especially true for CBRN weapons, which may not always be easily acquired by terrorist organizations and could potentially lead to a terrorist strike to obtain such resources. In order to implement machine learning of pattern recognition, several rigorous processes were utilized. Training and test data sets were determined and earlier data sets were used to predict new data, in an effort to develop a machine learning pattern recognition algorithm and model that could recognize patterns between past and future events.

Akin to pattern recognition, neural networks require the following components: (i) inputs, (ii) outputs, (iii) a training data set, and (iv) network topology. Specifically, the inputs are the selected core variables associated with an entry in a particular database (e.g. GTD or POICN) and the outputs are patterns recognized by the algorithm based on the inputs and the manner in which the neural network is trained. The training data set is a subset of the database (GTD or POICN) from the earliest entries, which is then used to generate output that can be compared and matched with test data set entries for validation of the algorithm. A subset of the pattern recognition algorithm developed can be run to aid in the identification of if and when specific terrorist groups adopt for instance CBRN weapons and moving forward which schemes are most probable in posing a security threat to society at large, and thereby enabling the intelligence and homeland security communities to make informed decisions regarding deployment of counter-terrorism resources to thwart terrorist plots before they occur.

Goals

1. Relate the five main categories of the POICN database:
   - Event, Agent, Target, Organization and Acquisition.
2. Determine any *related events, similar events* or *events by classifiers* and develop data sets for training neural networks.
3. Develop a Neural Network data set that will be used by the neural network, through Palisade and Matlab toolboxes, to make predictions on some of the potential missing data (e.g. -99 entries) of the POICN Database based on specific scenarios.
4. Measure the *Impact Value and Correlation* that some variables might have on others.

5. Determine potential scenarios for prediction:
    - Does Event A lead to Event B?
    - What is the relationship between organization parameters and the events that take place in relation to said organization?
    - Do their attacks grow bolder, more frequent as their numbers increase?
    - Is there a relationship between the values in the longitudinal tabs and the occurrence of the events?
    - Is something being easier to make in later years (e.g due to technology advancement) influence the use of that agent when compared to earlier years?
    - Determine potential patterns related to date intervals between attack events.
    - What about intervals between plot/interdiction/attack dates?

Methodology

In order to accomplish these goals, the Summer Research Team worked on determining relevant variables for each database (GTD and POICN) and carried out Machine Learning analysis on the inputs and targets related to each database. In the current timeframe, the emphasis of the GTD machine learning analysis was focused on Pattern Recognition Neural Networks for determining terrorist organizations related to a particular event and Regression Neural Networks for determining the correlation between the set of variables selected.

Knowledge Discovery in Databases (KDD) is a technique employed to find useful information from a database. The goal of KDD in databases is to convert the goals and requirements of the end-user into a “data-mining” goal. In our case, the first step was to determine which database variables to use for the neural network analysis. Subsequently, we determined potential scenarios for which particular variables might be useful, excluding those with a high number of unknowns. We carried out neural network tests to determine the effectiveness of specific variables on the misclassification percentages.

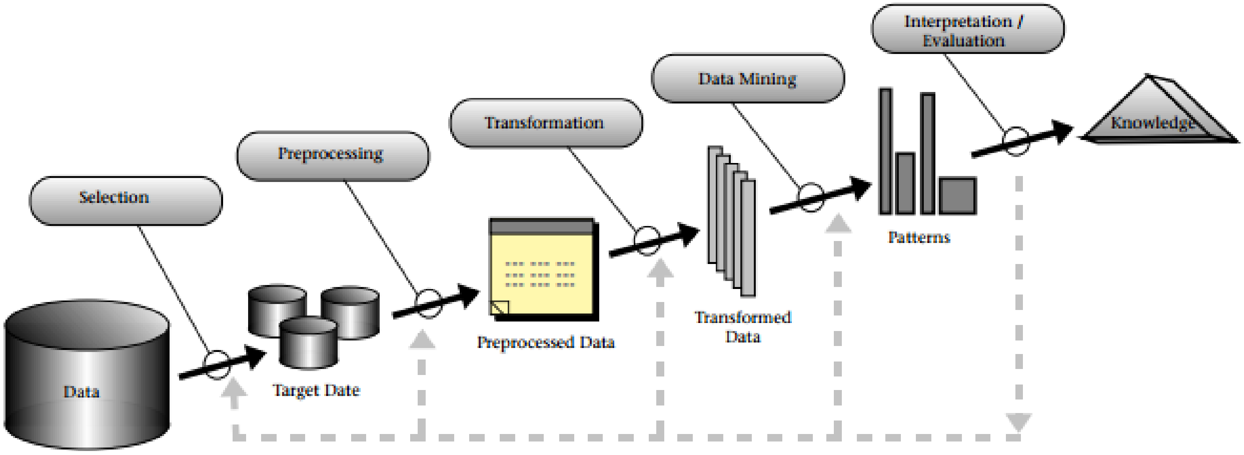
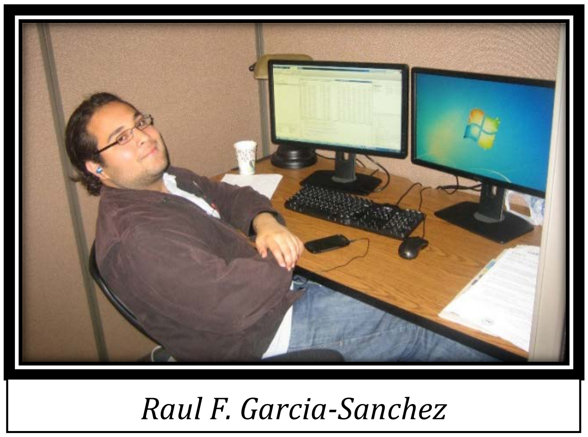


Figure 1. Process of Knowledge Discovery in Databases [1]

Global Terrorism Database (GTD)

A key concept in developing neural networks revolves around separating the data into three sets of data: (1) training, (2) test, and (3) prediction sets. The training set data is used to train the neural network and takes up most (70-80%) of the total data. The training data is further divided by the neural network into training, test and validation sets, which are used to determine when training results cannot be further optimized. Once the neural network is generated through training, the neural network is applied to the test set data in order to assess the network's performance, a process called cross-validation. The entire neural network structure flow can be seen in Figure 2.



Raul F. Garcia-Sanchez

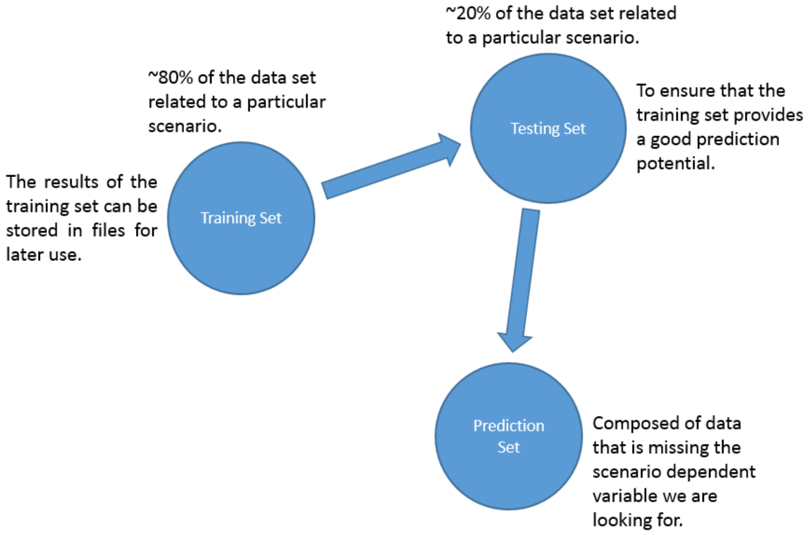

~20% of the data set related to a particular scenario.
~80% of the data set related to a particular scenario.
Testing Set
To ensure that the training set provides a good prediction potential.
The results of the training set can be stored in files for later use.
Training Set
Prediction Set
Composed of data that is missing the scenario dependent variable we are looking for.

Figure 2. The Neural Network Dataset Structural Flow Diagram.

Two particular datasets were employed for the GTD in order to find patterns in the data that would lead to recognition of the unknown terrorist organizations. The definition of these variables can be found in the codebook [2]. A snapshot of dataset 1 is presented in Figure 3.

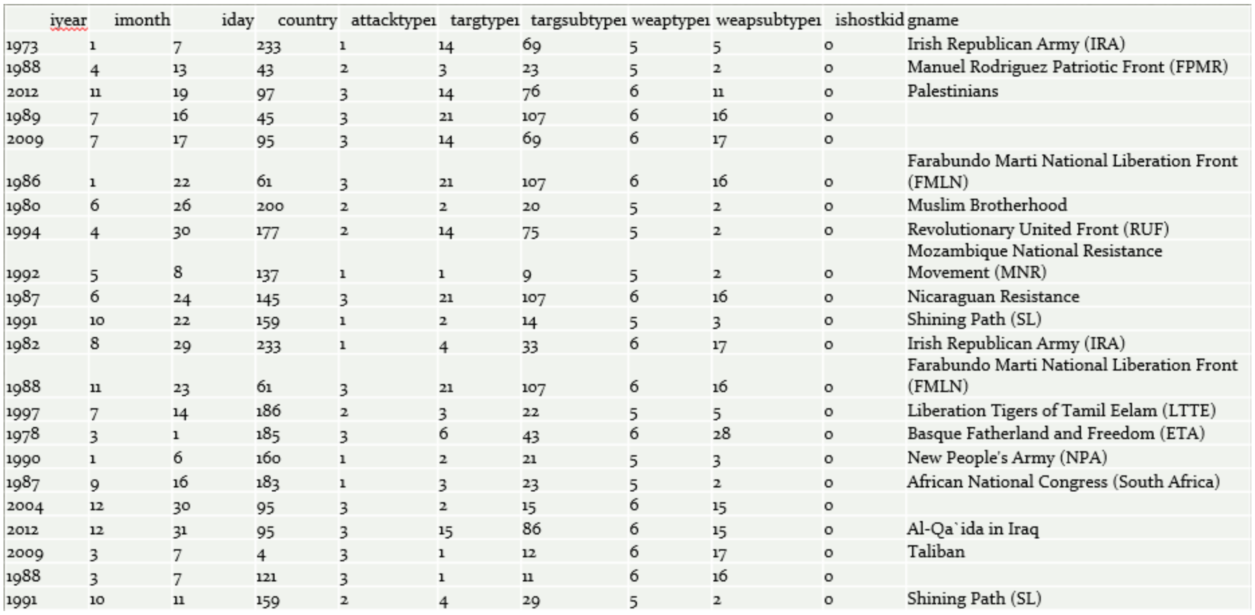
Dataset 1:

- **Inputs:** iyear, imonth, iday, country, attacktype1, targtype1, targsubtype1, weaptype1, weapsubtype1, ishostkid
- **Target:** gname

Some variations of this dataset included: (1) no subtype variables, (2), no subtypes or ishostkid.

Dataset 2:

- **Inputs:** iyear, imonth, iday, extended, country, region, vicinity, crit1, crit2, crit3, doubtterr, multiple, success, suicide, attacktype1, targtype1, targsubtype1, natlty1, guncertain1, weaptype1, weapsubtype1, property, ishostkid, INT_LOG, INT_IDEO, INT_MISC, INT_ANY
- **Target:** gname



iyear	imonth	iday	country	attacktype1	targtype1	targsubtype1	weaptype1	weapsubtype1	ishostkid	gname
1973	1	7	233	1	14	69	5	5	0	Irish Republican Army (IRA)
1988	4	13	43	2	3	23	5	2	0	Manuel Rodriguez Patriotic Front (FPMR)
2012	11	19	97	3	14	76	6	11	0	Palestinians
1989	7	16	45	3	21	107	6	16	0	
2009	7	17	95	3	14	69	6	17	0	
1986	1	22	61	3	21	107	6	16	0	Farabundo Marti National Liberation Front (FMLN)
1980	6	26	200	2	2	20	5	2	0	Muslim Brotherhood
1994	4	30	177	2	14	75	5	2	0	Revolutionary United Front (RUF)
1992	5	8	137	1	1	9	5	2	0	Mozambique National Resistance Movement (MNR)
1987	6	24	145	3	21	107	6	16	0	Nicaraguan Resistance
1991	10	22	159	1	2	14	5	3	0	Shining Path (SL)
1982	8	29	233	1	4	33	6	17	0	Irish Republican Army (IRA)
1988	11	23	61	3	21	107	6	16	0	Farabundo Marti National Liberation Front (FMLN)
1997	7	14	186	2	3	22	5	5	0	Liberation Tigers of Tamil Eelam (LTTE)
1978	3	1	185	3	6	43	6	28	0	Basque Fatherland and Freedom (ETA)
1990	1	6	160	1	2	21	5	3	0	New People's Army (NPA)
1987	9	16	183	1	3	23	5	2	0	African National Congress (South Africa)
2004	12	30	95	3	2	15	6	15	0	
2012	12	31	95	3	15	86	6	15	0	Al-Qa`ida in Iraq
2009	3	7	4	3	1	12	6	17	0	Taliban
1988	3	7	121	3	1	11	6	16	0	
1991	10	11	159	2	4	29	5	2	0	Shining Path (SL)

Figure 3. Snapshot of Dataset 1 in Excel.

In addition to using the above datasets, we also attempted to consolidate the datasets in order to determine whether or not there would be an impact on the percentage of pattern recognition misclassifications. Table 1 shows the variable count total going from unconsolidated to consolidated.

Consolidated Dataset 1:

- **Inputs:** idecade, iquarter, iweek, region, attackcat1, targcat1, weapcat1, ishostkid
- **Target:** gname

We utilized *Matlab* with the *Statistics and Neural Network toolboxes* to carry out the machine learning analysis involving the algorithms mentioned above. A script was developed to automate the data conversion portion of the research; in particular, the organization of the data for each of the neural networks developed and the conversion of the categorical variables into nominal arrays and dummy indicator variables was accomplished.

Table 1. Dataset 1 variable count, unconsolidated (left) and consolidated (right).

Variables	**Dummy Count**
Year	42
Month	13
Day	32
Country	188
Attack Type	9
Target Type	20
Target SubType	108
Weapon Type	8
Weapon Subtype	28
Ishostkid	1
Total	**449**

Variables	**Dummy count**
Decade	5
Quarter	4
Week	6
Region	13
Attack Category	4
Target Category	6
Weapon Category	3
Total	**41**

The neural network script carried out the following commands:

- Import Training/Test Set and Prediction Set from Excel Files.
- Prepare the Neural Network Data from the imported data.
- Partition the Training/Test for cross-validation.
- Sections for each Neural Network or Machine Learning Model, encapsulated with an if(false), allows user to select desired model to use when running the script.
- Similar structure for all models:
    - Breakup the combined Training/Test Set into set partitions.
    - Use training set to generate a model.
    - Use network generated on test set, then compare the results with the actual correct value.
    - Once the above step is acceptable, use network with prediction set data.

Two specific types of neural networks were employed throughout this project using the above datasets, namely fitting neural networks (NNF) and pattern recognition neural networks (NNPR).

Table 2. Neural Networks developed during GTD research.

Network Name	**Network Type**	**Predictor Type**	**Comments**
NN1	Fit	Numeric	First test for variable correlation using dataset 1.
NN2	Fit	Categorical	Better than NN1.
NN3	Fit	Dummy Indicators	Faster than NN1 and NN2 while averaging the same fitting results as NN2.
NN4-Sub-1	Pattern Recognition	Dummy Indicators	First attempt at pattern recognition using dataset 1.
NN4-Nosub-1	Pattern Recognition	Dummy Indicators	Removed subtype variables.
NN4-Nosub-2	Pattern Recognition	Dummy Indicators	Removed subtype variables and ishostkid.
Consolidated -Nosub-1	Pattern Recognition	Dummy Indicators	All variables consolidated as seen in Table 1.
Consolidated -Nosub-2	Pattern Recognition	Dummy Indicators	Country and Date unconsolidated, other variables consolidated.
MoreVars-1	Pattern Recognition	Numeric	Consolidation attempts high misclassification error suggests more variable approach using dataset 2. Additionally, test to see whether numeric or categorical are better in Pattern Recognition.
MoreVars-2	Pattern Recognition	Dummy Indicators	Consolidation attempts high misclassification error suggests more variable approach using dataset 2. Additionally, test to see whether numeric or categorical.

Through multiple neural network training iterations (Table 2), we were able to determine the impact of some of the variables on the misclassification percentages. Of particular note, we were able to determine that the subtype variables have no significant effect across networks that utilize the same source data. Additionally, preliminary results seem to suggest that data consolidation, at least the one attempted, significantly increases the misclassification percentage.

Results for regression and pattern recognition for each of the networks utilized can be seen in Figures 4-6 and Tables 3-5. All neural network results presented here were accomplished using Scaled-Conjugate Gradient (SCG) as the neural network training function.

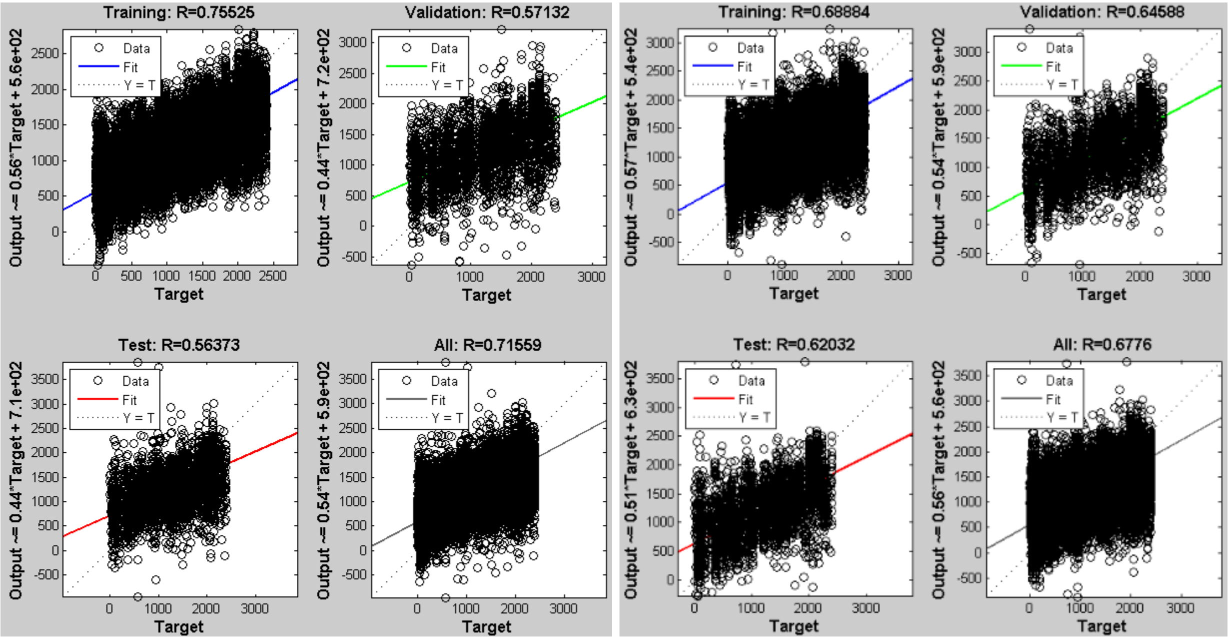


Figure 4. Regression Fit for NN2 (left) and NN3 (right)

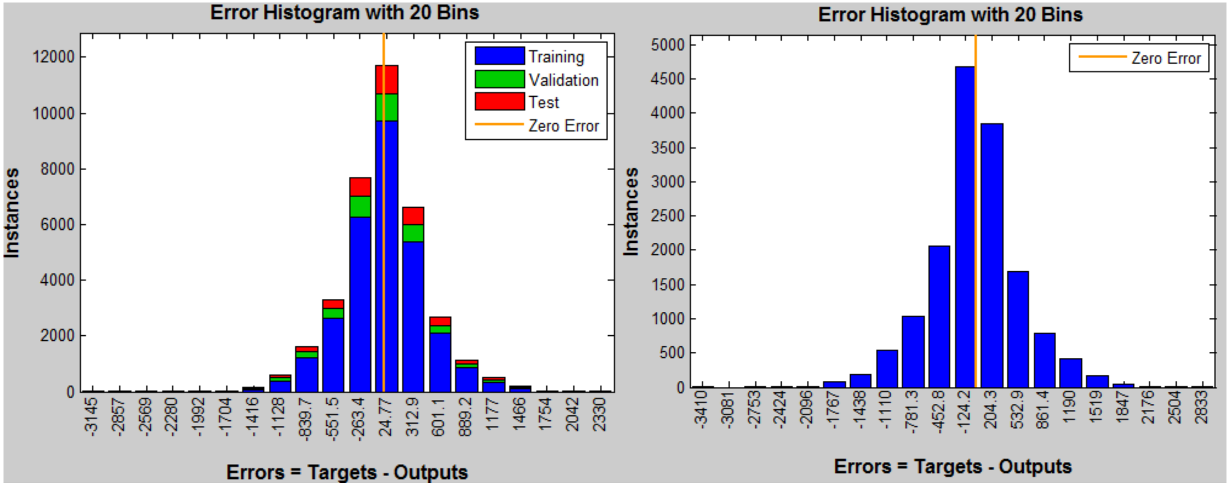


Figure 5. Error Histogram for NN2.

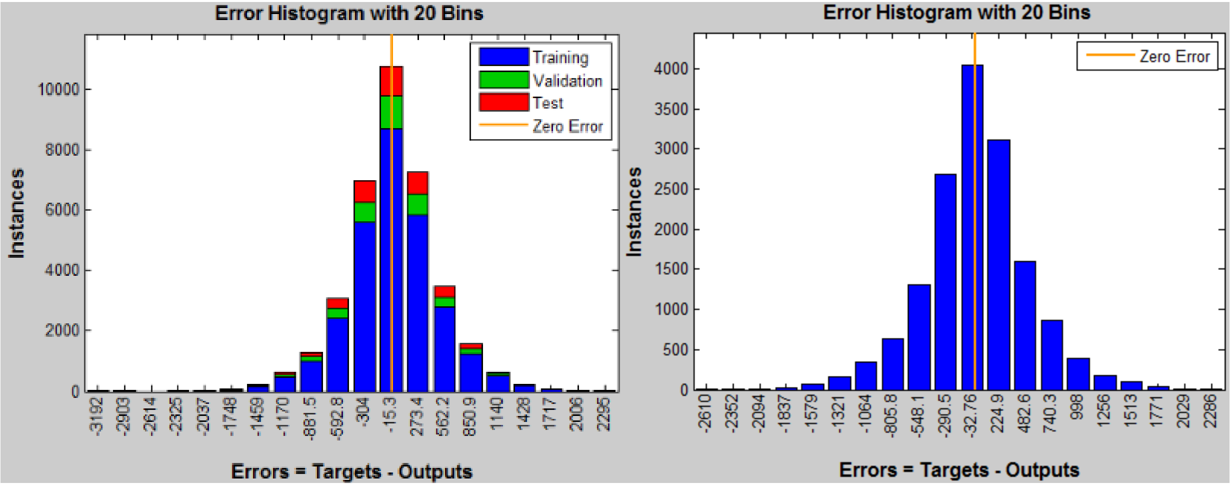


Figure 6. Error Histogram for NN3.

Table 3. Fitted Neural Network Regression Values for all Sets

Network Name	**Training Regression**	**Test Regression**	**Validation Regression**	**All Regression**	**Cross-validation Regression**
NN2	0.75525	0.56373	0.57132	0.71559	0.54862
NN3	0.6884	0.64588	0.62032	0.6776	0.64472

Table 4. Pattern Recognition Neural Networks employed and their misclassification error percentages (confusion).

Network Name	**Training Confusion**	**Test Confusion**	**Validation Confusion**	**Cross-validation Confusion**
NN4-Sub-1	21.84%	35.76%	34.00%	34.94%
NN4-Nosub-1	18.25%	32.85%	33.64%	32.55%
NN4-Nosub-2	19.47%	33.65%	33.40%	33.22%
Consolidated-Nosub-1	51.46%	59.06%	58.76%	58.72%
Consolidated-Nosub-2	22.53%	34.39%	34.63%	33.70%
MoreVarsNum	30.08%	37.09%	36.29%	36.38%
MoreVarsText	17.56%	30.22%	30.54%	29.77%

Table 5. Confusion differential between networks with and without subtype variables.

Network	**Training Confusion**	**Test Confusion**	**Validation Confusion**	**Cross-validation Confusion**
NN4Sub1	21.84327	35.76541	34.0033	34.94378
NN4NoSub1	18.24689	32.84691	33.64537	32.5538
Differential	3.59638	2.9185	0.35793	2.38998

Contributions made to project

- Scripting:
  - Data management script (table importing, data management, etc) (*GTDModel.m*)
  - Neural Network Data Manipulation (*GTDModel.m*)
  - Neural Network Training and Performance/Confusion (*NN4PatternScript.m*)
- Developed multiple neural networks of different types and parameters and tested their performance.
- Used neural networks (regression and pattern recognition) on GTD datasets.
- Consolidated low occurrence variable values and analyzed the behavior.
- Determined variables with low impact on misclassification percent.

New Skills and Knowledge gained

- Matlab Programming
  - Conversion of variables into dummy indicator variables
- Machine Learning
  - Clustering
  - Neural Networks

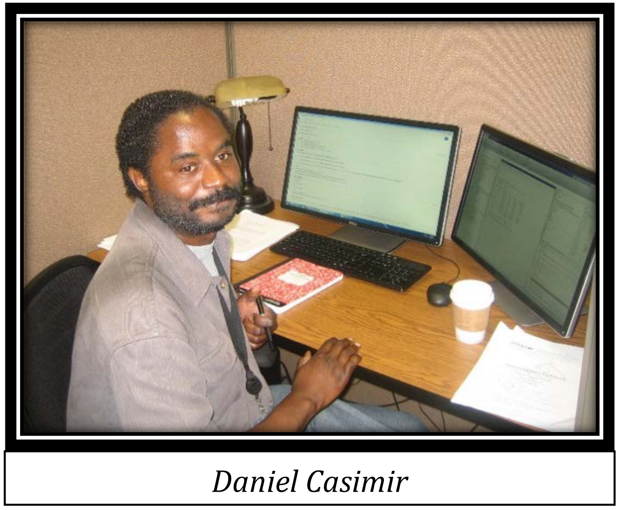
  - Combining machine learning algorithms

- Big Data
  - Data Mining
  - Managing
  - Consolidation
- Cloud and Cluster Parallel Computing
- Palisade software suite

The relevance of your research to the DHS mission

- Use of Matlab scripting and neural networks training to determine terrorist groups for future events added to the databases.
- Utilize the neural networks on scenario data - filling out the variables in a neural network dataset in a specific way in order to see which terrorism organization can be attributed to the scenario event.
- Modify script focused on GTD/POICN to match other database data and create neural networks based on that data.
- Eliminate variables that do not have a significant impact on the misclassification error when training neural networks based on terrorism-related databases.
- Determine the correlation between selected variables and the way machine learning algorithms separate them into clusters.

Profiles of Incidents involving CBRN by Non-state actors (POICN) database

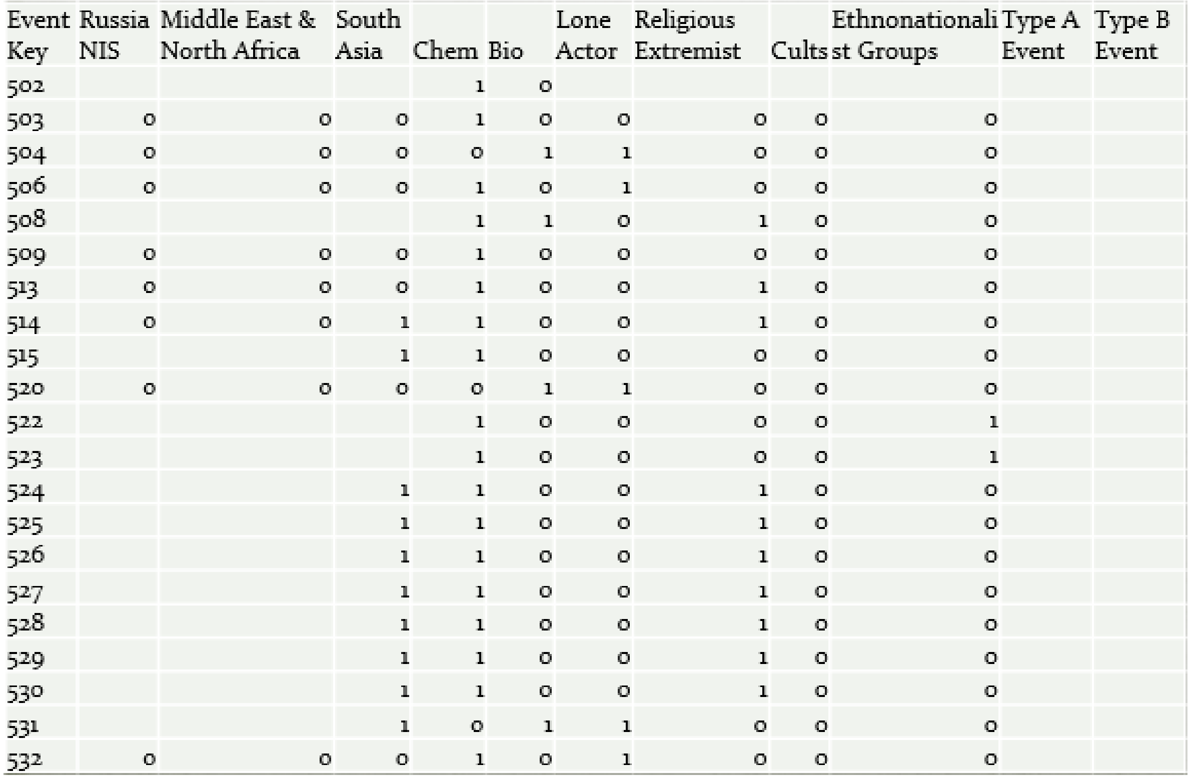


Daniel Casimir

The growing potential for CBRN attacks by terrorist organizations is now firmly established. The POICN database available at http://www.start.umd.edu/, includes terrorist activities (plots, acquisitions, weaponization, and attacks) of non-state actors either seeking or utilizing CBRN agents. The POICN database is recognized as one of the foremost comprehensive open source databases on CBRN terrorist events that occurred from the early 1990's up to 2011 [3].

One way the relational POICN database differentiates itself from other similar terrorist databases is by classifying source validity and the inclusion of variables that rate the uncertainty inherent within and between sources of event information [4]. Multiple variables included in the source evaluation scheme are aimed at capturing the distortion of information regarding CBRN activities that is in-built in the data set [5-6]. For example, each source was coded for its *competence*, and its *objectivity* with respect to each event. Also, the *credibility* variable in increasing order of reliability (Level 1, 2 or 3, respectively) is coded once for each event in the database and is based on corroboration between several independent sources.

The POICN database demarcates eight (8) different types of CBRN events, namely: use of agent, attempted use, and threat with possession, acquisition of a weapon, acquisition of an agent, attempted acquisition, plot, and protoplot. The eight categories are separated into two broad classes: Type A (seeking a CBRN weapon) and Type B (possessing a CBRN weapon). Both types A and B are important and in order to differentiate between the two, an approach is to use the data mining tool of *logistic regression* analysis, one of the numerous data mining tools in the emerging field of Knowledge Discovery through Databases (KDD).



Event Key	Russia NIS	Middle East & North Africa	South Asia	Chem	Bio	Lone Actor	Religious Extremist	Cults	Ethnonationalist Groups	Type A Event	Type B Event
502				1	0						
503	0	0	0	1	0	0	0	0	0		
504	0	0	0	0	1	1	0	0	0		
506	0	0	0	1	0	1	0	0	0		
508				1	1	0	1	0	0		
509	0	0	0	1	0	0	0	0	0		
513	0	0	0	1	0	0	1	0	0		
514	0	0	1	1	0	0	1	0	0		
515			1	1	0	0	0	0	0		
520	0	0	0	0	1	1	0	0	0		
522				1	0	0	0	0	1		
523				1	0	0	0	0	1		
524			1	1	0	0	1	0	0		
525			1	1	0	0	1	0	0		
526			1	1	0	0	1	0	0		
527			1	1	0	0	1	0	0		
528			1	1	0	0	1	0	0		
529			1	1	0	0	1	0	0		
530			1	1	0	0	1	0	0		
531			1	0	1	1	0	0	0		
532	0	0	0	1	0	1	0	0	0		

Figure 7. Snapshot of a sample POICN Dataset.

Figure 7 shows a snapshot of the POICN dataset. Figure 8 shows the proportions among the 165 sample events that we selected from the 471 total events in the most recent version of the POICN database that had credibility ratings of 1, 2, or 3 that was used in our logistic regression model of event type prediction. Table 6 provides the corresponding count and percentage breakdown of the POICN credibility levels of the events shown in Figure 8.

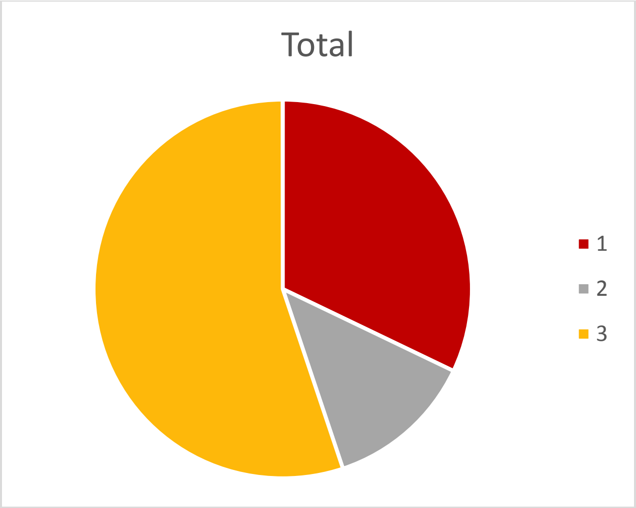


Figure 8. Pie chart breakdown of Credibility Levels of POICN events

Table 6. Breakdown of Credibility levels of POICN events

Credibility level	**Count of CREDIBILITY**	
	Count	**Percent**
1	53	32%
2	21	13%
3	91	55%
Grand Total	**165**	**100%**

This portion of the summer project was a continuation of work being done by [3-4], and our progress so far in our initial reproduction of ref. [4] results using current POICN data has been promising. We were able to run a “logit” or log-odds ratio model with no interactions using the same 9 binary predictor independent variables that were used in ref [4], and arrived at similar conclusions regarding the impact of each variable on a sample CBRN event being either of type A or type B, for all but two of the variables among the sample data, namely the variable RussiaNIS used for events occurring in Russia or one of the new independent states and the variable Bio used for events where biological weapons were mentioned.

A comparison of the actual results, showing the values and signs of the coefficients of both our and Breiger et al's simple logit model are shown in Table 7. In the future it is planned to continue reproducing the results of Breiger et al.'s novel use of the logistic regression technique, and finally to apply the more recent data mining tool of neural networks to this same problem of discovering CBRN event type and compare the results from the two approaches, and looking at the pros and cons of both methods.

Table 7. Coefficients for the logistic regression model variables of Breiger's and our model.

Labels	Breiger et al's coefficient estimates	Our coefficient estimates
Intercept	-0.25	-4.0533
RussiaNIS	2.2	-0.084233
Middle East	1.35	0.30323
South Asia	3.06	0.26297
Bio	-0.24	0.33975
Chem	1.9579	0.86233
Lone Actor	-0.56	-0.52259
Religious Extremist	-2.58	-0.25766
Cults	-2.49	-0.30323
Ethno nationalists	-2.06	-0.16353

Contributions made to the project

- Identified and extracted the highest credible sample events from the POICN database to use in an initial logistic regression analysis of CBRN event type prediction.
- Organized the sample data from the POICN database and ran an initial linear logistic model regression analysis using Matlab's Machine Learning based Generalized Linear Model tools.

- Reproduced the conclusions reached by Breiger et al. for most (7 out of 9) of the binary variables used in the logistic regression modeling of the prediction of CBRN event type (Type A, or Type B) based on current data from the POICN database.

New skills and knowledge gained

- Introduction to the steps and data mining methods used in the emerging, interdisciplinary field of Knowledge Discovery through Databases (KDD).
- Exposure to and use of the supervised machine learning method of linear regression analysis, as implemented in Matlab.
- Introductory lessons in the Farsi language, through participation in one of the numerous enrichment programs provided throughout the internship, at UMD's START center.

Relevance of your project to the DHS mission

- This project falls in line with the START center's mission of KDD and other data-driven methods in an effort to gain an understanding of the "human causes and consequences of terrorism…" This project was specifically focused on terrorism events involving weapons of mass destruction, whether they are chemical, biological, radiological, or nuclear (CBRN).

Presentations made to START Staff

Four presentations were scheduled and delivered (jointly by P. Misra, R. Garcia-Sanchez and D. Casimir) to START personnel as part of the summer research experience and are summarized below:

(1) Demonstration of an Actual Neural Network Data Set from POICN 2.2 (*June 10, 2014*);

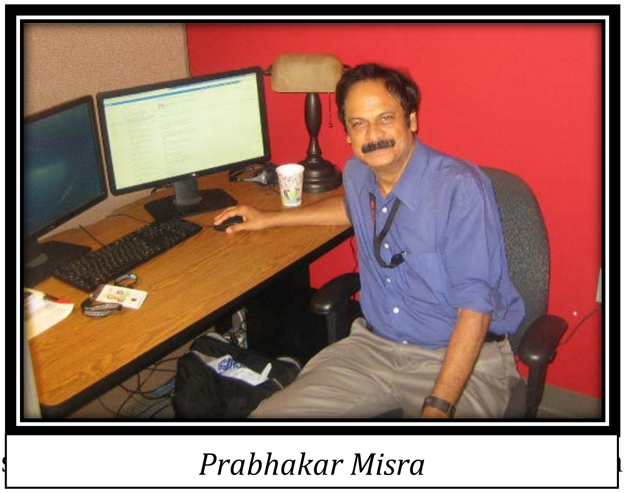
(2) Neural Network Model Development for Terrorist Group Prediction (*July 1, 2014*);

(3) Pattern Recognition Neural Network Development for Terrorism Event Databases (*July 23, 2014*); and

(4) Machine Learning Algorithms for POICN and GTD (*August 5, 2014*).

The graduate students, Daniel Casimir and Raul Garcia-Sanchez, gave an additional research presentation to the summer interns at START on August 8, 2014, titled "Innovative Algorithm and Database Development Relevant to Counterterrorism and Homeland Security Efforts at START".

Impact of Research Experience on Academic and/or Career Planning



Prabhakar Misra

Involvement with this DHS summer research project enabled all of us to learn new techniques and tools related to handling and working with "Big Databases", such as GTD and POICN at START. We became familiar with some powerful software packages *Palisade* and *Matlab* and the associated toolboxes. The diagnos variety of useful fitting techniques, namely how to review the large data set and plan the analyses, create a StatTools Data Set, perform single-variable and multi-variable analyses, examine relationships between variables, create a logic model, and perform regression and risk analyses. In *Matlab*, the various toolboxes provided a wealth of learning related to the statistical analyses associated with large data sets. Specifically, for instance the Neural Network Toolbox in *Matlab*, facilitated training in the following important areas: function approximation and nonlinear regression, pattern recognition and classification, clustering, time series and dynamic systems, and neural network control systems and architectures.

The experience gained with these software packages will be directly applicable to statistical analyses of physics data gathered in laboratory experiments and in coursework problem sets and assignments, both at the undergraduate and graduate levels.

Future Work

There are some issues that remain to be ironed out with respect to the script (as opposed to using the neural network toolbox interface in Matlab). While the original approach of using the Matlab interface provided a good starting point, we found that the time for the training process increased steadily as the data subsets and variables grew in complexity. We have made some initial cloud computing runs using the Amazon EC2 resource with 16 CPUs/node and 244GB RAM/node, which has improved our training steps by at least a factor of 6 (i.e. 20 hours versus 120 hours) and documented via the neural network figures showing the time to completion. We are currently also working on getting Matlab commands input into the University of Maryland's Deepthought2 HPCC supercomputer. Due to the extremely large size of our datasets, the neural network training has taken a significant portion of resources and time. We expect that once the batch scripting files necessary to run Matlab on Deepthought2 are developed that we will be able to speed up the neural network process considerably. Additionally, other neural network training techniques can be explored to test their effectiveness against the Scaled-Conjugate Gradient (SCG) methodology.

An alternate approach would be to implement Fast Fourier Transform (FFT) techniques for enhanced efficiency in processing large data sets, however initially we wanted to test out how the neural networks would behave with the raw GTD and POICN data. Running FFT on the current data, carrying out other mathematical and book-keeping operations on the same data and running it through the neural network process, would be worth exploring.

On the POICN front, we have been in contact with Prof. Ronald Breiger of University of Arizona whose earlier work on logistic regression [4] complements our neural network approach and we plan on further interactions and possible collaborations with him and other START personnel as we move forward in refining and streamlining our neural network based algorithms and models for handling large data sets.

Classroom Plan

The DHS Summer Research Experience provided valuable training in developing mathematical models relating to large data sets and performing high-level statistical analyses. Prof. Misra is scheduled to teach a graduate level course during the AY 2014-15 where these statistical tools and techniques will prove very useful. The two-semester course Statistical Mechanics I and II (PHYS 222 and PHYS 223, 3 Credits each) is designed to cover the following topics: Ensemble Theory, Classical and Quantum Statistics, Dense Gases and Liquids, Magnetism, Applications in Solid State Physics, Kinetic Theories, and other cutting-edge Special Research Topics, such as Laser Trapping & Cooling and Bose-Einstein Condensation). The modeling and statistical analyses tools learnt and used during the summer research experience will be directly applicable to the two courses during the Fall 2014 and Spring 2015 semesters in the homework and other assignments that would be required deliverables from the students taking the course. For example, following development of a linear regression model for a given data set, the students can be assigned to do the following interpretive tasks: load sample data and define predictor variables, fit the linear regression model, perform analysis of variance (ANOVA) for the model, decompose ANOVA table for model terms, display coefficient confidence intervals, and perform hypothesis tests on generated coefficients. Detailed interpretation of the linear regression results will involve clear understanding of the following: (i) degrees of freedom and the coefficient estimates for each corresponding term in the model, (ii) standard error of the coefficients, (iii) t-statistic for each coefficient, (iv) root mean squared error, (v) subtle difference between R-squared and adjusted R-squared coefficients, (vi) F-statistic to understand the relationship between the response variable and the predictor variables, and (vii) the meaning of the p-value for the F-test on the model. For a given model, it will also be important to display and interpret confidence intervals for the regression coefficients, since confidence intervals provide a measure of precision for the linear regression coefficient estimates.

For practical purposes, a neural network is a large interlocked collection of simple elements or so-called neurons, which are two-state (on-off) devices that switch from one state to another [7-9]. The insights gained by applying the technique of neural networks to large data sets (e.g. GTD and POICN) for pattern recognition will come in handy in understanding the properties of disordered systems via the study of cooperative behavior associated with large, highly connected networks of neuron-like processors in some physical systems, such as strongly coupled nonequilibrium systems. Physicists have studied neural networks by drawing an analogy between such networks and magnetic systems (e.g. spin glasses) [7]. Spin glasses exhibit haphazardly distributed ferromagnetic and antiferromagnetic interactions, and the low-temperature phase of such systems is an archetype example for condensation in chaotic systems. Such systems can be studied using the techniques of statistical mechanics in the second-half of the course offering in Spring 2015.

Acknowledgments

The authors of this report, Prabhakar Misra, Raul Garcia-Sanchez and Daniel Casimir, would like to thank the Department of Homeland Security (DHS) for funding through the Summer Research Team Program for Minority Serving Institutions that enabled them to work with mentors (Gary Ackerman, Markus Binder, John Sawyer, William Braniff and Eamon Diab) at *START – National Consortium for the Study of Terrorism and Responses to Terrorism* on this exciting research project *Innovative Algorithm and Database Development Relevant to Counterterrorism and Homeland Security Efforts* as applied to the Global Terrorism Database (GTD) and the Profiles of Incidents involving CBRN by Non-state actors (POICN) efforts at START.

References

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, Fall 1996, pp. 37-54.

[2] GTD Codebook. Available at: http://www.start.umd.edu/gtd/downloads/Codebook.pdf. Accessed: Aug-06-14.

[3] Ronald L. Breiger, Gary A. Ackerman, Victor Asal, David Melamed, H. Brinton Milward, R. Karl Rethemeyer, and Eric Schoon, "Application of a Profile Similarity Methodology for Identifying Terrorist Groups That Use or Pursue CBRN Weapons", Book Chapter: Springer-Verlag, (2011), pp.26-33 Lecture Notes in Computer Science, Volume 6589 Social Computing, Behavioral-Cultural Modeling and Prediction. 4th International Conference, SBP 2011, College Park, MD, USA, March 29-31 2011. Proceedings Editors: John Salerno, Shanchieh Jay Yang, Dana Nau, and Sun-Ki Chai.

[4] Ronald Breiger, Paul Murray, Lauren Pinson, "Patterns of CBRN Use by Non-State Actors: Analyzing the Evidence", [Prepared for and presented at the Annual Convention of the International Studies Association (ISA), San Francisco, April 4, 2013. Panel on New Data for the Scientific Study of Conflict.]

[5] Gary A. Ackerman and Lauren Pinson. 2011. "Speaking Truth to Sources: Introducing a Method for the Quantitative Evaluation of Open-Sources in Event Data." College Park, MD: National Consortium for the Study of Terrorism and Responses to Terrorism (START Center): working paper.

[6] John Sawyer and Gary Ackerman. 2012. "Promethean Journeys: Examining the Mechanisms by Which Terrorists Acquire New Technologies of Lethality." Paper presented at the Annual Meeting of the International Studies Association, San, Diego.

[7] H. Sompolinsky, “Statistical Mechanics of Neural Networks,” *Physics Today*, December 1988, pp. 70-80.

[8] J.J. Hopfield and D.W. Tank, “Computing with Neural Circuits: A Model,” *Science* 8 August 1986, pp. 625-633.

[9] J.J. Hopfield and David W. Tank, “Computing with Neural Networks,” *Science* 6 March 1987, pp. 1228-1229.