Rare Events

Contact: D. McMorrow - dmcorrow@mitre.org

October 2009

JSR-09-108

Approved for Public Release

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 983-6997

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
October 2009	Technical	

4. TITLE AND SUBTITLE

Rare Events

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
13099022

5e. TASK NUMBER
PS

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
JASON Program Office
7515 Colshire Drive
McLean, Virginia 22102

8. PERFORMING ORGANIZATION REPORT NUMBER

JSR-09-108

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OSD
3030 Defense Pentagon
Washington, DC

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

JASON was asked by the Department of Defense (DoD) to conduct an evaluation of the nation's ability to anticipate and assess the risk of rare events. "Rare events" specifically refers to catastrophic terrorist events, including the use of a weapon of mass destruction or other high-profile attacks, where there is sparse (or no) historical record from which to develop predictive models based on past statistics.

This study was requested by the Strategic Multi-Layer Assessment (SMA) program, which is part of the Joint Staff/J-3, STRATCOM/GISC, and the Rapid Technology Program Office within the Department of Defense Research and Engineering.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Hriar Cabayan
Unclassfied	Unclassified	Unclassified	UL		**19b. TELEPHONE NUMBER** *(include area code)* 703-746-1453

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Contents

1 INTRODUCTION

JASON was asked by the Department of Defense (DoD) to conduct an evaluation of the nation's ability to anticipate and assess the risk of rare events. "Rare events" specifically refers to catastrophic terrorist events, including the use of a weapon of mass destruction or other high-profile attacks, where there is sparse (or no) historical record from which to develop predictive models based on past statistics.

1.1 Study Motivation

This study was requested by the Strategic Multi-Layer Assessment (SMA) program, which is part of the Joint Staff/J-3, STRATCOM/GISC, and the Rapid Technology Program Office within the Department of Defense Research and Engineering. The SMA program was established in 1997 in response to the need for multi-agency and multi-disciplinary approaches to support the ten US Combatant Commands in complex operations outside their core competencies. One such area of complex operations that cuts across the Commands' expertise is dealing with the Weapons of Mass Destruction/Terrorism (WMD-T) threat space.

The SMA undertook an effort in June of 2007, at the request of United States Special Operations Command and United States Strategic Command, to develop the foundation for establishing a sustainable, federated intelligence community (IC) wide WMD-T intelligence and operations analysis enterprise [3]. The goals set forth to SMA for the operations analysis enterprise are the following:

- Anticipate how terrorists are likely to acquire and use WMDs over the next ten years.
- Provide means to target areas, entities and persons facilitating adversary WMD courses of action.
- Characterize the global WMD-T environment.
- Identify and name areas, entities and individuals of WMD-T interest.
- Identify and prioritize WMD terrorist courses of action.
- Identify and prioritize collection requirements.

Specifically, the SMA was charged with the development of scientifically sound theory and methodology, leading to a collaborative infrastructure to accomplish those goals. This effort is on-going and has led to the request for the JASON study.

1.2 JASON Study Charge

SMA recognizes that current practices rely on decision options generated by a limited number of people with varying degrees of expertise, often with access to classified information. Empirical predictive models are beginning to be proposed for use to anticipate rare events, but the predictive accuracy of these models is unknown. SMA has begun to engage with the academic research community in developing predictive modeling approaches, and has begun to develop collaborative experiments using strategic gaming methods such as an upcoming Limited Operational Exercise [1].

[1] A Limited Operational Exercise is a multiplayer experiment designed to exploit and study information sharing and collaboration[35].

JASON was asked to evaluate the following:

1. Can it be done?
    - Can one use scientific models to accurately anticipate the existence and characterization of WMD-T threats?
    - What current models are being used and how good are they?
    - What metrics can one use to assess accuracy?
    - Are studies in more measurable/quantifiable domains applicable to threat event characterization?
2. Are collaborative experiments useful in this domain?
    - Do experiments using collaborative processes test and improve predictions of future rare event threats?
    - Is there value in bringing outside experts into the process to generate actionable decision options?
    - What should one hope to get out of these collaborative experiments?
    - How does one measure success of the collaborations?
3. Is academic expertise important in real time decision options formulation?
    - If having academic experts involved in a decision option model is desired, how would one do it?
    - Would this add noise or insight?

Early anticipation and amelioration of "rare events" raise questions that are fundamentally about human behavior. This is the research domain of

Table 1: Study Briefers

Hriar Cabayan (POC)	OSD
Gary Ackerman	START
Victor Asal	University of Albany
Chris Bronk	Rice University
Krishna Pattipati	University of Connecticut
Paul Whitney	PNL
Robert Popp	NCI
David Lazer	Harvard
Mike Stouder	University of Michigan
Tom Rieger	Gallup
Dan Flynn	DNI
Susan Numrich	IDA
Elisa Bienestock	NSI
Joan McIntyre	ODNI
Fred Ambrose	USG

the social sciences. Special emphasis in the study will be placed on the role of social science methods and expertise.

1.3 Briefers and Materials Reviewed

JASON was introduced to the problem by the briefers in the following table. Materials recommended by the briefers, along with a wide range of other classified and publically available materials were reviewed and discussed by JASON. These included three key papers solicited by SMA [19], [21], [20] and an extensive RAND report [26] on the topic.

Throughout our briefings is was stressed that countering WMD-T is the joint responsibility of many agencies and organizations. Ideally, a federated IC-Wide WMD-T intelligence and operations community should be able to coordinate and integrate intelligence, define the operational environments,

describe the impact of the operational environments, evaluate adversaries, and determine adversaries' potential courses of action.

In current thinking, WMD-T threats are often conceptualized as a combination of *intent*, *capability*, and *opportunity*, where all three are required for a WMD-T rare event to occur, so denying any one would prevent the event. Another common conceptualization is in terms of *where*, *who*, *what*, *why*, *when*, and *how* – locations of interest, groups or individuals of interest, weapons they may use, their motivation, their operational timing, and their process – where understanding these variables could help strategic planning and allow for the development of tactics, training, and procedures [2].

The slide in Figure 1, presented to JASON [4], depicts the WMD-T "Operational Spectrum" and represents the focus of SMA's work. The goal is to provide actionable intelligence to decision makers so threats can be accurately anticipated and characterized "far left of boom." Ultimately, the nation needs a systematic method to prevent threats from evolving into actions and if that fails, accurately attributing the actions to the correct sources. This may require new degree of cooperation between intelligence analysts, outside experts, and law enforcement.

1.4 Summary of the Study

The "rare event" of interest is an extreme, deliberate act of violence, destruction or socioeconomic disruption, such as an attack of 9/11 scale or greater. It is not a realistic goal to anticipate and prevent all rare events, but it may be possible to make rare events rarer, and to reduce their effect.

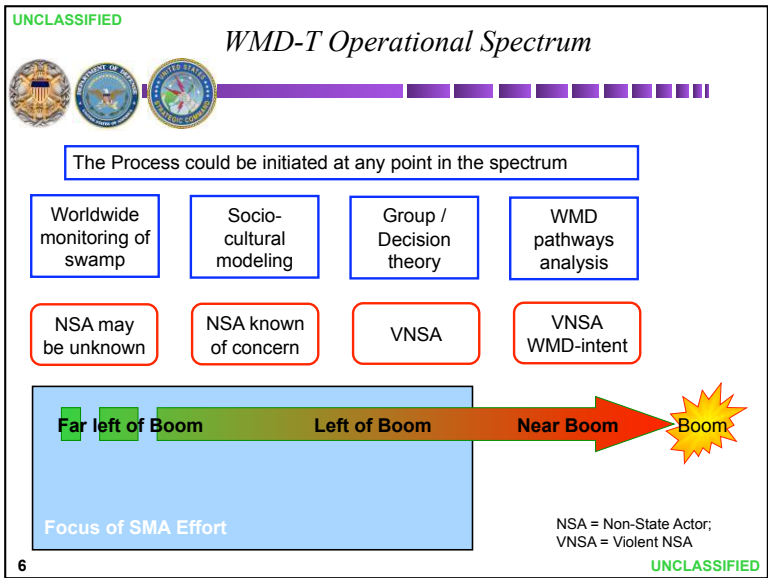


Figure 1: WMD-T Operational Spectrum denoting the focus of attention for SMA.

A rare event is preceded by a chain of individually more likely developments that create intent, capability, and opportunity. Intervention may be possible at many points in that chain – ranging from social policy decisions that reduce the probability of radicalizing groups of people who may seek violent means to express their grievances, to tactical intelligence and law enforcement that detects an imminent event and arrests would-be perpetrators. If possible, quantitative analysis and mathematical models to objectively optimize intervention decisions should be used.

There are two principal problems in applying quantitative models to the anticipation of rare events. One problem is that rare events are rare. There will necessarily be little or no previous data from which to extrapolate future expectations in any quantitatively reliable sense, or to evaluate any model. In the extreme, how can the probability of an event that has never been seen or may never even have been imagined be predicted? The second problem is that the mechanisms at work are largely human behaviors, not just physical

forces. This leads to the hope for quantitative social sciences models (e.g., of causes), not only physical models (e.g., of physical consequences and damage). Is it feasible to build useful social sciences models that can inform our decision-making, given not only that the behavior of humans is immensely complicated and unpredictable to begin with, but also that human adversaries (unlike physical disasters) will react and adapt to our planning to try to make ineffective our plans to thwart them?

1.5 Overarching Conclusions

Predicting WMD-T rare events is obviously a desirable goal. This has been stated many times by DoD, DHS, the IC, academic researchers, and others. However, it is simply not possible to validate (evaluate) predictive models of rare events that have not occurred, and unvalidated models cannot be relied upon. An additional difficulty is that rare event assessment is largely a question of human behavior, in the domain of the social sciences, and predictive social sciences models pose even greater challenges than predictive models in the physical sciences. Reliable models for ameliorating rare events will need to address smaller, well-defined, testable pieces of the larger problem.

JASON recognizes the problem put forth is exceedingly hard. This report will describe some possible paths forward. First, we give a concise response to the three primary points in the Study Charge. This will be followed with our overarching findings and recommendations.

1. **Can it be done?** There is no credible approach that has been documented to date to accurately anticipate the existence and characterization of WMD-T threats. Experience from the natural sciences and engineering provide guidelines for how to characterize certain aspects of the risks involved, but are of limited value or applicability at the present time. Social science approaches pursued to date are far less well developed, and not even at the point at which their utility can be evaluated, as currently applied. No reliable metrics of accuracy have yet been identified, and there is a significant deficiency in applying standard approaches from engineering and science such as false alarm rates and signal detection in the face of massive clutter.

2. **Are collaborative experiments useful in this domain?** Collaborative experiments are of limited value because they are based on an as-yet-unproven assumption that lack of communication and collaboration is the key choke point in anticipating WMD-T threats. No clear objectives or metrics have, so far, been identified for collaborative experiments.

3. **Is academic expertise important in real time decision making?** Area expertise and real-world experience appear to be highly valuable in addressing the problem at hand. Some of this expertise is available in academia, but not exclusively so. There is no evidence that academics necessarily have better (or worse) capability in this regard.

The combined urgency of the rare event threat, the difficulty of evaluating rare event models, and the complexity of social sciences problems has led some to advocate the suspension of normal standards of scientific hypothesis testing, in order to press models quickly into operational service. While

appreciating the urgency, JASON believes such advice to be misguided. The threat of "rare events" will be with us for a long time. Like finding a cure for cancer or predicting earthquakes, this is a difficult research area that will most likely make progress in many small steps. Experience in the development of many other scientific fields shows the importance of adhering to rigorous scientific standards, so that small successes are tested, communicated, critically examined, reproduced, and built upon; thus a field as a whole gains steady and lasting traction, even though near-term actionable progress may seem elusive. Although patient husbandry of a long-term research program may fall short of addressing the immediate operational needs, JASON believes it is the best way forward for success in the long term.

Findings:

- There is a clear need to establish a solid, rigorous, long-term foundation for applied research, development, and operations. This research will be rooted in social sciences and involve academics.
- There is danger in premature model building and the use of such models, to the exclusion of careful data collection.

Recommendations:

- Evaluate how DoD program choices meet the best standards and practices in empirical research across all of science. This includes defining baselines and measures of success.
- Clearly express the purpose of each activity in terms of the problems being addressed.

- Collect and share data within and between agencies, and with the academic community.

In this report, we expand on these general recommendations with the following series of discussions.

Section 2: We discuss "motivation" as a useful addition to the intent/capability/opportunity conceptualization, especially because some of the most impressive social sciences research in our briefings and readings was more relevant to motivation than intent.

Section 3: We discuss different types of scientific models, distinguishing predictive models from "insight" models that build intuition, and distinguishing point prediction models from risk assessment models. We emphasize examples of risk assessment models as an area from other fields in which the probability of large rare events is usefully extrapolated from small common events.

Section 4: We discuss some rules for evaluating the rigor of predictive models and insight models, from a high level of scientific practice. We emphasize the availability of primary datasets as an important foundation for any field, enabling others to reproduce and build on one's findings.

Section 5: We discuss the importance of characterizing false positive prediction rates (false alarms) in detecting rare events.

Section 6: We discuss the strengths and limitations of strategic games to generate insight, including red team exercises and the SMA's planned Limited Operational Exercise. We describe a game theoretic view of antiterrorism defense.

Section 7: In the final section, we summarize JASON's own past work on one type of rare event scenario: bioterrorism.

Throughout, we frame most of our discussion of WMD-T threat assessment in terms of related problems in the biological and physical sciences. How are analogous "rare events" problems analyzed for natural catastrophes, such as earthquake, wildfires, or weather forecasting? How are predictions made and risks evaluated for rare events? When do physical scientists start to develop models rather than emphasizing empirical primary data and intuition? What different sorts of models do they build? How do they evaluate whether these models are useful? From this standpoint, we discuss our view of the current state of the art in applying quantitative models, frequently from the social sciences, to "rare event" anticipation,[2] and we discuss ways to think about directing and evaluating a portfolio of research investments in this area.

[2]We frequently encountered references to "Black Swans" in our study. The Black Swan metaphor was popularized by a recent book *The Black Swan: The Impact of the Highly Improbable* by Nassim Taleb [39]. The metaphor has clearly had great impact on how people are thinking about rare events, so we considered Taleb's argument carefully. This discussion is included in Appendix A.

2 MOTIVATION: "FAR LEFT OF BOOM"

Understanding *intent* has been proposed as a strategy to move "far left of boom" in the anticipation of WMD-T rare events (see Figure 1). Specifically what is meant by understanding and then measuring intent is not obvious. The white paper *From the Mind to the Feet: Assessing the Perception-to-Intent-to-Action Dynamic* [21] argues there is a lack of robust theories and valid measures of intent. The definition of intent is given as "a determination to act in a certain way for a certain purpose, a mental construct." Part of the ambiguity may be due to the fact that this conflates two different concepts: "motive" (the purpose) and "intent" (the determination to act).

Perhaps there is a lesson to be learned from law enforcement and the law, where motive and intent are distinguished. A grandson may be significantly in debt, thus giving him motive to kill his rich grandmother. Until the grandson actually decides to kill the grandmother, he does not have intent. With this distinction in mind, much of the white paper ([21]) and of the materials presented from the briefers in Table 1 actually has more to do with motive than intent.

One reason the distinction is important is that specific intent is not only difficult to determine, but also may be too "close to boom" to be useful for strategic planning. Intent is more of a tactical intelligence question. Motive, on the other hand, may be more measurable and actionable and be a more suitable framework in which to think about "far left of boom" planning and interdiction.

Finding: Focusing on motivation as opposed to intent in the context of WMD-T reduces the need to anticipate specific events, and is better suited to a strategic rather than tactical approach to reducing rare event probabilities.

Recommendations: Add "motive" to the conceptual framework of intent, capability, and opportunity.

2.1 Lessons about Motivation versus Intent

In this section we build a taxonomy for the understanding of motivation(s) that could lead to WMD-T threats. This taxonomy, given in Figure 2, should be useful for the organization and planning of data collection and modeling efforts.

In *law enforcement*, evaluation of the potential for "rational actor" criminal activity is based on motivation, intent, target vulnerability, and guardian capability [22]. Suppose we assume constant motivation for terrorists to inflect harm. The limit for carrying out intentions driven by constant motivation will be the accessibility of desirable targets and the vigilance of those protecting the targets. Indicators of terrorist activity, in this case, will be comparable to those of criminal activity. This is because terrorists will have similar operational needs, including weapons acquisition, financing, false documents, sanctuary, support. Pre-empting terrorist action then can be based on target prediction, monitoring, and disruption of operations components. The consequences of constant motivation move across the bottom of Figure 2 and up the center.

Using a *deterrence perspective*, the law enforcement procedures can be augmented by understanding the nature of the non-constant motivation. Two limiting cases distinguishing intent (actions) are those of actors driven by strategic motivation, and those driven by internal logic [23]. Strategic motivation is basically the desire to effect global political goals, and to do so by manipulating adversarial political and social entities. For many terrorists this would equate to imposing their worldview on others. In contrast, the motivations of internal logic are driven by desired affects with the actors' own political and social entity. In this case the internal goals may require external actions, but the external impact is not an end in itself. For terrorists this would equate to developing power within their own organization, with internal status and recruitment of new members enhanced by external actions. It also couples with externally frustrating actions directed at preventing resolution of external conflict.

Using the deterrence perspective, prediction and pre-emption tactics can be expanded beyond those based on a constant motivation model. Targets can be differentiated by their strategic versus internal value. The intrinsic value of strategic targets makes protecting them an effective way of frustrating strategic actors. In contrast it is much more difficult to predict or protect the broad range of targets that may be attractive to an internal-logic actor. The deterrence perspective moves across the top of the taxonomy in Figure 2 and down the middle.

A further useful distinction, based on a social science perspective, would differentiate terrorist motivation in terms of ideologically-driven actors (as in the deterrence perspective) versus actors driven by immediate social/material needs. The non-ideological social/material actors (foot soldiers) can represent a pool of external supporters or potential recruits to an ideologically-

driven group. Positively addressing their immediate needs, and thus reducing their affinity to terrorist organizations, is a mechanism of deterrence. Understanding their perspective is a method of understanding target choice by the ideologically-driven actors who wish to maintain their support. This *social* component is shown in Figure 2 to feed the both law enforcement (constant motivation) and deterrence (non-constant motivation) characterizations of motivations.

Finally, it is important to understand the range of choices that are made in assessing motivation and intent [24]. Analytical communities, charged with rigorous assessment, tend to focus on capabilities analysis in which the development of capability is equated to intent to use the capability. Decision makers, those who must allocate resources to address the problem, are more likely to focus on behavioral signals. These can run the gamut from negative (e.g. suicide bombings) to positive (e.g. agreeing to discussions). Thus for assessment programs to be effective in driving policy decisions, they must address the full range of technical to behavioral aspects of terrorist activities. This important concept is captured in Figure 2 along the right hand side.

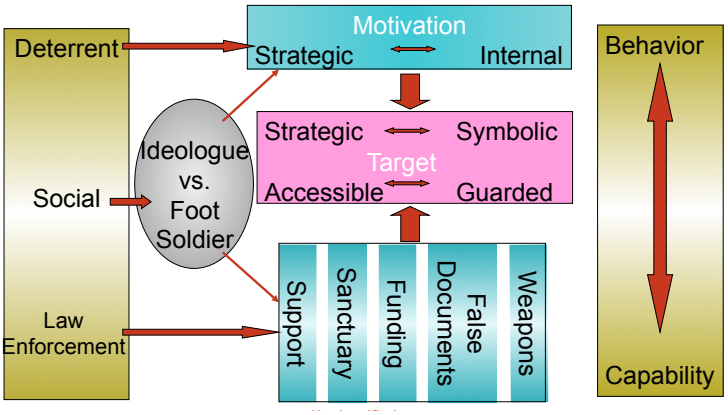


Figure 2: JASON Taxonomy for Measuring Motivation

2.2 Detecting Sociological Motivations for Terrorism

One method of determining motivations for terrorism is to conduct opinion polls that query attitudes of people in regions that have been a source of terrorist activity. There are many organizations that poll world opinion. The Pew Charitable Trust [16], World Opinion [14], and National Opinion Research Center [17] conduct global surveys whose results are publically available. The START Center[3] has conduced several polls across the Islamic world [44, 45, 46]. Private researchers also conduct polls that include questions that can be relevant for understanding attitudes toward terrorism. Gallup ([15], [25]) has identified regions likely to become more unstable within the next five years and have given clear confidence levels for their estimates. Expert analysis of these polls may provide actionable information for governments as well as providing questions that might be desirable to pose in future polls.

The opportunity to reduce terrorist attacks by changing motivation is recognized by many governments around the world, and many governments have vigorous deradicalization efforts [40, 41, 43]. There are programs that work with people who have been detained in connection with terrorism investigations. The programs then use some combination of doctrinal revision and rewards to induce the people to abandon violence. Other programs attempt to dissuade people from terrorism using programs that are not targeted directly at individuals but address groups of people, where the groups generally include people who are not at present members of terrorist organizations [42].

[3]START is the National Consortuim for the Study of Terrorism and Response to Terrorism, located at the University of Maryland (http://www.start.umd.edu).

These programs assume that certain attitudes are linked with a propensity for terrorism.

2.3 Removing Opportunity and Capability

For an attack to occur motivation, intent, and capability must combine with opportunity. It is possible to thwart an attack by removing any one of these elements. Opportunity can be reduced by physical security. Governments around the world are taking vigorous action to improve physical security, where the security measures are based on assessments of previous attacks and estimations of who is likely to commit the attack, where the attack is likely to occur, and how it is to be carried out. It is certain that these assessments do not include all possible devastating attacks. Section 6.3 gives a theoretical view of the allocation of physical security resources.

However excellent the existing physical security, it is still desirable to develop techniques for avoiding attacks that do not involve our ability to predict any details of the attacks in advance. Similarly, capability can be reduced by controlling access to technical knowledge and materials, if one assumes that we correctly envision the nature of those attacks. However, there are many attacks for which capability barriers have become extremely low, as in the case of biothreats (see Section 7).

Reducing the threat of a devastating attack by removing opportunity or capability relies on the government having a good picture of the nature of the attack. Predicting details of rare events can be very challenging, even in a static environment where past data could be a good basis for future behavior. Unfortunately, terrorist threats are dynamic and evolve precisely to counter security measures. Predictions based on the past may lead to erroneous

conclusions that lead to large resources being expended in ways that do not reduce the threat of attacks. Even if the government has correct vision of a possible attack, prevention through removing opportunity or capability requires that the government control access to people, materials and venues required for the attack. Except for certain special cases, such as a nuclear attack, it is not possible for the government to control the access required for that attack. Thus, it is desirable to search for a counterterrorism approach that is not dependent on accurately predicting future attacks.

Eliminating the motivation removes all possible threats without requiring any knowledge of the space of possible attacks and does not require that the government control access, to people, places or things that might be used in an attack. Eliminating the motivation for attacks offers a tremendous advantage; however, this approach presents a challenge since is assumes that there is a link between something that the government can manipulate and the behavior of terrorists.

3 MODELS

In all areas of study, many different kinds of models are used. It is important to distinguish the use of models for providing subjective experts with *insight* into a hard problem, versus the use of models that aim to make objective *predictions*. For predictive models, there is a distinction between models for *probabilistic risk assessment* on long time scales (the probability that some event will happen in the next ten years, for example) versus *specific point prediction* of individual rare events. Placing these topics in the WMD-T threat assessment context leads to the following findings and recommendations.

Findings:

- Social science-based models do not yet exist for anticipating and interceding in rare WMD-T events.
- It is unreasonable to aim for predictive models of specific rare events.
- Predicting human behavior and evaluating any predictive models of rare events, predicated on human behavior, are difficult; however, prediction of signatures of concern may be possible.

Recommendations:

- Frame issues such as training, operational planning, monitoring, and mitigation in forms that allow (social science) models to tackle limited goals, with well defined questions, and testable hypotheses.

- Define how the model will help in decreasing the probability or impact of a rare event.

3.1 Point Prediction Models of Rare Events

For rare events, by definition it will be impossible to validate any model that seeks to produce specific point predictions of a rare event. This is because by "rare event" we are talking about any event that hasn't happened yet and will only happen once. Even if we had a correct model, we won't know it's valid until it's too late to intervene in the event.

There may be one way forward for predictive modeling of rare events. Suppose we assume rare events are drawn from a distribution that includes *other* events that are sufficiently common that we can observe many of them, enough to evaluate a model. We could, for instance, assume that small observable events and large rare events are sufficiently related in their causes that we are willing to assume large rare events are just the high-magnitude tail of some underlying distribution of events. For example, imagine a (social science) model that predicts a terrorist cell with actionable accuracy, without trying to predict what event the cell might attempt to precipitate. We might be able to find ways to evaluate such a model if we had access to enough of the right sort of data on actual terrorist cells. We could imagine testing interventions that reduce the number of terrorist cells, the probability of events perpetrated by terrorist cells, or shift the event magnitude distribution towards smaller events we can count and measure.

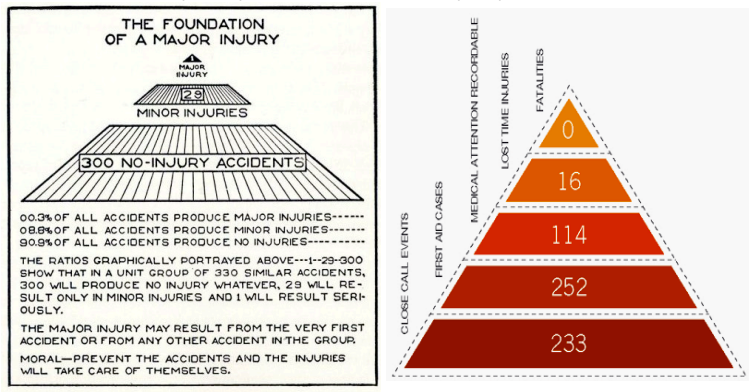
Such an approach sharply focuses one's attention on the key assumption: that the large number of countable small events share enough causal similarity to a rare event that predicting small events corresponds to pre-

dictability of large rare events. This assumption may be invalid – and if it were, this would be impossible to know, for the same reasons that a model of rare events alone cannot be validated. Nonetheless this suggests a way forward for the expectations one might demand from predictive social science models. First, a model should predict "small" events that occur with sufficient frequency that the model's predictions can be evaluated (validated). Second, the model should explicitly relate the frequency of smaller to larger events, so it can extrapolate to an unobserved tail of high-magnitude events. Third, the assumption that the high-magnitude tail of rare events is drawn from essentially the same distribution should be carefully considered. Such predictive models would require good, large datasets of events and incident data that enable event prediction.

3.2 An Example: Frequency/Magnitude Distributions

The approach described above is in fact commonly used as an approach to rare-event problems in other fields. The area of safety has used this concept since the 1950s, creating the safety pyramid, as given in Figure 3. This community has found causal relationships leading to the conclusion that small accidents lead to medium accidents lead to catastrophic accidents. Eliminating the base of the triangle has become a best practice strategy for safety.

When the available data on event frequency versus magnitude shows a simple relationship over a wide range of magnitudes, (e.g., if ten-fold larger events are systematically a hundred-fold less likely), then we might feel reasonably comfortable with extrapolating the tail of the distribution into predicting the probability of higher magnitude events beyond any events observed thus far. This is not a point prediction model. It does not predict



H.W. Heinrich (1950), *Industrial Accident Prevention, 3rd edition.*

2003 Data, Pinnacle West Corporation *http://www.pinnaclewest.com*

Figure 3: The Safety Pyramid

any single specific event at a specific time. Rather, it provides a quantitative prediction of the *probability* that an event of a given magnitude will happen within some (long) time interval. Unlike predicting individual rare events, such a model can be evaluated for observable events, because it predicts that future observed events will continue to follow the same frequency/magnitude distribution. This prediction can be used to help guide decisions about deployment of resources for risk management and attenuation.

One good example of this type of modeling is earthquake risk assessment. If all that is known is that large earthquakes are possible, we might worry that a "rare event" earthquake of enormous magnitude might occur at any moment and utterly destroy an American city. How could we possibly plan for such an unknowable catastrophe? Quantitative prediction of specific earthquake events is not yet achievable. However, earthquake magnitudes are observed to follow a distribution called the Gutenberg-Richter law, where ten-fold larger earthquakes on the Richter scale (Richter 8 versus 7, or 7 versus 6), are empirically observed to occur with ten-fold smaller frequency. This probability distribution is also called a *power law distribution* (also known as a *Pareto distribution*, or *Zipf's law*). Figure 4, taken from

[30], demonstrates this relationship for earthquakes in southern California.

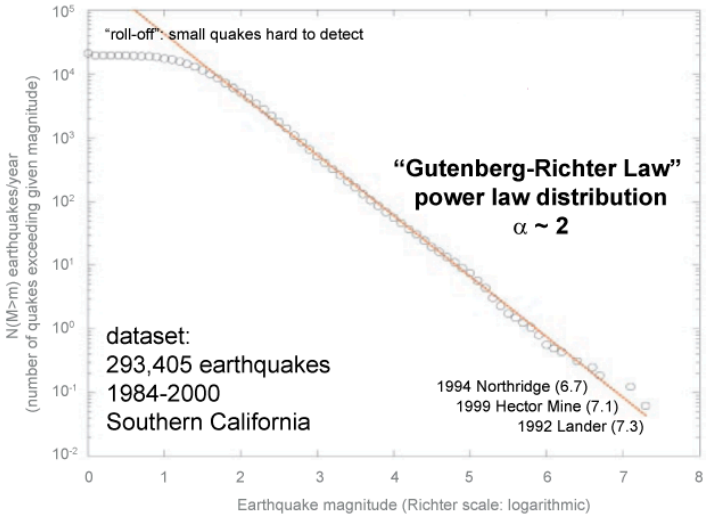


Figure 4: Power Law for Southern California Earthquakes

Empirically observed frequency/magnitude distributions are one of several factors used for quantitative risk assessment for different locations. The US Geological Survey publishes earthquake risk assessments based in part upon such distributions [29]. Building codes, insurance companies, and other earthquake disaster preparedness and risk mitigation efforts take these predictions into account. (For an excellent introduction to power law distributions and several ways in which they may arise, see Newman [18].) A power law is a "long-tailed" distribution, which means that rare high-magnitude events will occur with far higher frequency than a "normal" (Gaussian) distribution would predict.

The model has caveats. First, it is a curve-fitting exercise, not a mechanistic physical model of earthquake geology. It is not entirely apparent why the physics of earthquakes leads to this empirical distribution, and this should induce distrust. Are the physics such that one really expects to obtain this same distribution from every individual fault? Indeed, this is an

ongoing argument in the seismology community. Second, can we reliably extrapolate the high magnitude tail into unobserved rare events? Clearly there is a finite limit on earthquake energy, so the extrapolation must break down. Third, are high-magnitude earthquakes caused by the same underlying physical processes (drawn from the same distribution) as low-magnitude earthquakes? Seismologists are divided on this issue. Fourth, at some point we might want to worry that there is another unsuspected process that can also produce a "rare event" earthquake, a process with extremely low frequency but off scale magnitude when it occurs, e.g., the seismic shock of a huge asteroid strike.

3.3 Power Laws and Rare Events

What relevance does this sort of probabilistic risk assessment have for predicting "rare events" having to do with terrorism? Consider the following reasonably well defined prediction question: what is the probability that a terrorist event of larger scale than 9/11 (in terms of civilian fatalities) will occur somewhere in the world in the next decade?

The frequency/magnitude distribution of terrorist events has been studied by Clauset, *et al.* [31]. Figure 5 shows the power law fit for terrorist events between 1968 and 2006. The observed data empirically appear to follow a power-law distribution. Clauset et al. [31] fit a power law of $\alpha = 2.38$ to these data, indicating that events of $10x$ greater magnitude in terms of deaths are about 240-fold less likely ($10^{-2.38}$-fold). Is the 9/11 attack in New York City (2749 killed)[4], despite being an order of magnitude more catas-

[4]The MIPT database defines an "event" as a single target in a single city in a single day, so 9/11 is recorded as three "events": New York, Washington DC, and Shanksville Pennsylvania.

trophic in loss of life than any other large terrorist events (such as the 2004 Beslan school hostage crisis that killed about 400) an outlier on this fitted relationship? Perhaps surprisingly, according to the power law distribution in Figure 5 it was not. The probability of an event of 9/11 magnitude or greater in the 1968-2006 period is about 23%, according to the fitted distribution, (see Appendix B for this calculation).

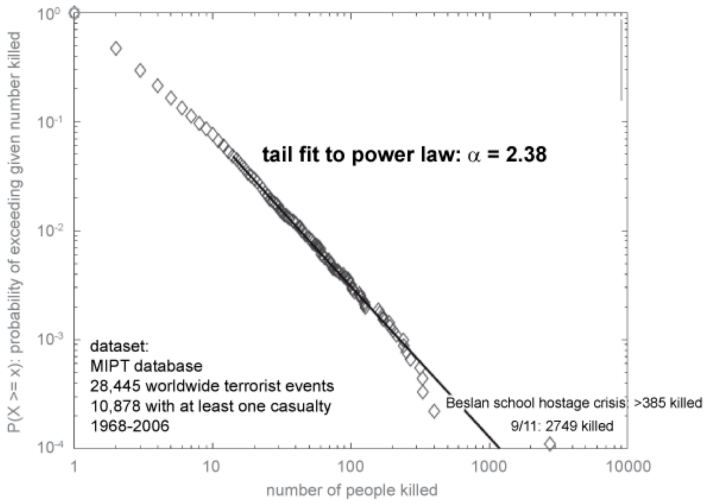


Figure 5: Power Law Fit to Deaths from Terrorist Events. Redrawn from Figure 2 of [31].

A more detailed look at these data are give in Figure 6, from the same study [31], where plots for injuries and total casualties are included. In this figure, the tail of the observed data distribution even more clearly encompasses the 9/11 event in New York.

We haven't quite answered the original question yet. The power law fit for the data in Figure 5, shows the magnitude *per event*, not the number of events *per time*. If we believe this fit and want to predict not only the magnitude of the next event larger than 9/11 but also its probability of occurrence in the next ten years, we need to factor in the frequency of events. The data in [31]

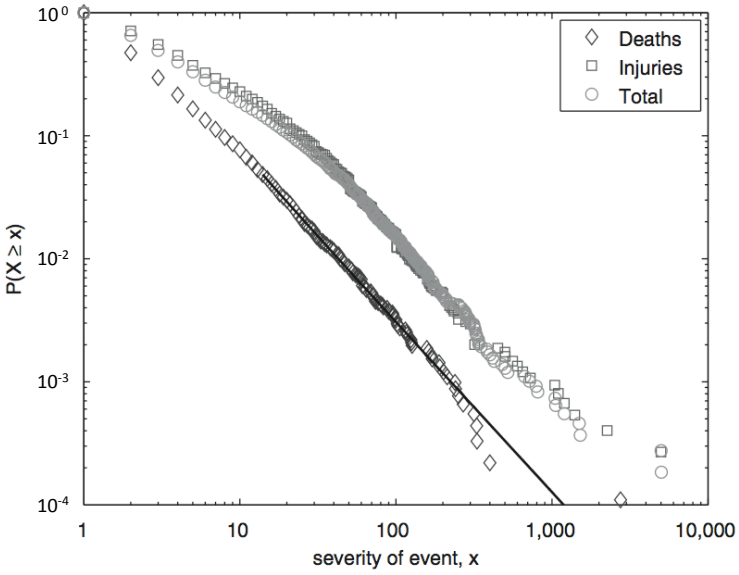


Figure 6: Injuries and Deaths from Terrorist Events. Redrawn from Figure 2 of [31].

include 28445 recorded terrorist events (10878 of which killed or injured at least one person, and 9101 of which killed at least one person), over a time span of 38.5 years (1968 to mid 2006). If we assume that event frequency is constant with time (more on this naive assumption later), this means about 240 events per year, or 2400 events per decade. Clauset *et al.* [31] doesn't quite give enough information about their data or their power law fit to reproduce their work exactly, but we can approximately deduce their fit from the figure. With the estimate of events per decade and the power law fit in Figure 5, we calculate that another 9/11-scale event in the world is unlikely but not improbable in the next ten years. It has a probability of about 7%, (see Appendix B for this calculation).

Now, let's consider this prediction carefully. On the positive side, this example shows that it certainly is possible to create a well-defined, empirical, data-driven prediction of a rare terrorist event. It is the case that we can make empirical statements about the probability of events so extreme that

none have yet been observed, provided we assume rare events occur on a continuum with more frequent events. On the other hand, how much faith should we put in this prediction? We can use this almost trivial example (the "model" has just three free parameters; two for the power law tail fit, one for the event frequency per year) as a case study in how aggressively the assumptions and predictions of a quantitative model should be dissected.

Do we really believe that event frequency is constant with time? Probably not, although Clauset *et al.* [31] did look at sliding windows of two-year intervals and found relatively little temporal variation on that scale in the 1968-2007 interval. We should hope that the assumption is false, because according to this model, a good way to reduce the probability of a rare event is to reduce the frequency of all events.

Is there really just one frequency/magnitude distribution? Almost certainly not; surely there's really a very complex mixture of different distributions that happens to yield a smooth overall distribution. For example, surely a terrorist's choice of weapon must make a difference in the expected number of casualties. Clauset *et al.* [31] show some analysis of this, breaking their data down into explosives, firearms, blade weapons, and other types, and they do find different distributions. Thus our single empirical distribution is really a mixture of different distributions. That's all right, as long as each component has been observed. But now suppose that another component of our distribution tends to cause very high-magnitude events; is so rare that it has not been observed yet; and is nonetheless not as rare as the power law extrapolation suggests. An empirical distribution gives no insight into the contribution of any rare unobserved components, and we may greatly

underestimate the probability of this other very high-magnitude rare event. Terrorist use of a nuclear weapon might be such an example.

Is this information actionable? This kind of "prediction" enables probabilistic risk assessment, not point prediction of an specific event. Prediction of specific earthquake occurrences is not currently possible. A risk assessment model does not specifically tell us who, where, when, why, or how. Risk assessment models are most useful for prioritization of different risk scenarios in order to optimize resource investment. For example, earthquake risk assessment is used to assign different levels of risk to different geographic regions, and expensive earthquake-resistant building codes are implemented for high-risk areas. To be useful in the WMD-T rare events context, a risk assessment model would similarly need to be factored somehow into areas of differing risk and differing resource allocation; perhaps by geographic region, perhaps by some classification of terrorist group type, or perhaps into the type of event (biological, chemical, nuclear, or unconventional attacks). It must be noted that earthquake risk assessment models for different geographic regions do not rely solely on earthquake statistics, but also on seismic fault mapping and an understanding of the underlying physics; no such parallel exists for terrorist behavior.

One of the basic difficulties in rare events scenarios, as demonstrated above, is that we do not have enough data to fully sample the probability distributions governing terrorist actions. This applies to the geographic distribution of terrorist events and to the event size distribution. In the face of this problem, a standard idea is to appeal to Bayesian methods where *a priori* information is used to augment small samples. In the technical note in Appendix C, we point out that there has been some progress on using Bayesian

ideas to estimate the entropy of incompletely sampled distributions. The entropy is directly tied up with the notion of predictability; the purpose of the note, then, is to review these new ideas and suggest they be applied to the problem at hand.

3.4 Insight Models

Predictive models are not the only kind of models scientists use. *Insight* models are used to build expert intuition – such as visualizing complex data sets, or just helping to modularize and structure the steps in a mental model of a problem. Predictive mathematical modeling is the most scientifically demanding way in which models are used, but it is probably not the main use of models in science. The main use of models in science is to develop intuition for hard problems. Models are used to illustrate, visualize, and analyze a problem, to help human experts see patterns in data, and to systematize an expert's thinking in a way that might reveal key gaps in knowledge about the problem.

An insight model need not be complicated. A simple systematic cartoon on a napkin may suddenly reveal a missing facet of a problem. Other models may be complicated. A red-team exercise may reveal an unanticipated vulnerability; an agent-based simulation may help illustrate inefficiencies and bottlenecks in resource allocation; a social network analysis may help clearly visualize a pattern of connections between people in a large dataset.

Experts develop their own ways of organizing and viewing their data as they think about a problem - such as drawing cartoons showing relationships, or developing a personal system of archiving and indexing data. Experts develop these models for themselves, and they learn from the experience of

other experts in their field. Because experts spend most of their time doing their job rather than developing new tools, there is good reason to fund free-standing research and development projects into new (insight) models.

From a programmatic standpoint of funding research, the main problem with standalone research projects that aim to create new (insight) models is that they separate the model's creator from the model's user community, so they tend to face an adoption barrier. Experts are rightly skeptical of new tools developed by non-experts, especially if a model appears complex, mathematical, and highly abstracted rather than hewing closely to real-world data analysis needs. Success of an insight tool should ultimately be judged by how many experts use it and find it indispensable in their work. "Useful to experts" necessarily includes many factors that become just as important as the scientific validity of the model – issues such as software quality and usability, in the case of computer models. Therefore an important part of any research plan to develop new models is the researchers' plan for collaboration and adoption by experts. Will the tool be used and evaluated by real-world analysts? Do they find it useful? Will it spread to other analysts if it is successful?

Bioinformatics is an example of a field in which there is much research and development of quantitative/computational models for hard, ill-defined problems that may not be satisfactorily evaluated by their developers as objectively predictive, and where such models are instead best judged by how useful experimental biologists find them to be in suggesting new hypotheses and experiments. To prevent a natural tendency for model developers to cocoon themselves into an isolated artificial community away from real-world needs, NIH program officers demand concrete collaboration plans and evidence that such models will be evaluated by their successful use in the hands

of subject experts. For example, one recent NIH bioinformatics program announcement includes the following language:

> " Given the expanding needs in biomedical research for advances in a variety of areas of information science and technology, the approaches and technologies proposed under this announcement should ultimately be generalizable, scalable, extensible, and interoperable. The projects should take into account the needs of the biomedical research community that will be the ultimate end users of the products of the research. The projects should also address plans for ensuring the dissemination of useful products of the research, including approaches, technologies and tools, to the relevant research and user communities. The informatics and computational research proposed should be future-oriented, fill an area of need or projected need, and seek to exceed the current state of the art. [NIH Biomedical Information Science and Technology Initiative (BISTI) Program Announcement PAR-07-344 Innovations in Biomedical Computational Science and Technology [36]."

It is important to build a science base for the development and use of insight models. Based on several reviews of social science research for rare events ([26], [19], [21], [20]) there is little evidence the researchers build their work on the work of others. Several briefers from Table 1 commented that the sharing of data is strained. This is not an issue of collaboration. This is an issue of one person's insight or results feeding into the model and results of the next researcher. This low degree of cumulative development collective inhibits the establishment solid, accepted science base.

We conclude this subsection with the following finding and recommendations specific to insight models.

Finding: Insight models are useful, but in a different way than predictive models. Accordingly, they will be evaluated more subjectively, by how useful they are to experts

Recommendations:

- Require developers of insight models to demonstrate the value of those models by putting them in the hands of subject matter experts.
- Define how the model and knowledge gained from the modeling will add value to the cumulative, community, understanding of the problem.

4 MODEL EVALUATION AND DATA Model

This discussion of model evaluation (validation) is prompted by claims JASON encountered suggesting that theory in social sciences should be exempted from the usual tests of a scientific theory. The following, from a recent comprehensive report *Social Science for Counterterrorism* [26], are excellent examples of what we heard.

> " Another theme of our work is that, even where the social science is "strong", the paradigm of reliable prediction is usually inappropriate. Too many factors are at work, many with unknown values and some not even knowable in advance. Except in rare cases in which matters are over determined, there will be a substantial "random" component in social behavior."

> " The objective of analysis in social science should often not be reliable "prediction", but rather an understanding of possibilities and perhaps of rough probabilities or odds."

These statements are misleading because in fact, many scientific models only make predictions in terms of probabilities, yet we still demand that the predictions of such models be evaluated. This leads to the following findings and recommendations, applicable to both predictive and insight models.

Findings:

- Opportunities exist for development of formal quantitative and qualitative models for anticipating and interceding in rare events.
- Predicting human behavior and evaluating any predictive model of rare events is difficult.

Recommendations:

- Tackle limited goals, with well-defined questions, and testable hypotheses.
- Demand rigorous definition and evaluation of ALL models.

4.1 Model Evaluation

To apply a model in any meaningful way in science, including social science, one should try to meet the following three principles:

1. A useful model should attempt to encapsulate, in some satisfactory way, our scientific understanding of the underlying behavior.
2. The domain of applicability of any model must be specified.
3. The implicit assumptions of any model must be specified.

Some examples of unsatisfactory models which violates the first principle include models that violate any known physical laws or mathematical theorems, models that are under-determined or that contain fudge factors

that can be adjusted to give whatever result may be sought (rendering them meaningless), and models that invoke components outside the realm of science, such as inputs from religious belief systems.

All models are approximations. Otherwise excellent models that violate the second principle and continue to be used outside their domain of applicability tend to generate incorrect answers. Examples include Galilean relativity whenever velocities begin to approach those of the speed of light, or representing light by rays that travel in straight lines when the dimensions are either too small (because of diffraction) or too large (because of gravity).

The third principle reminds us that models are useful simplifications and are therefore based on simplifying assumptions. Some assumptions are so obvious that they occasionally go unstated, whereas other assumptions can be very subtle and therefore hard to recognize. Assumptions affect the second principle in a fundamental way: whenever the model assumptions are violated, the model is not applicable.

Models that claim to be objectively predictive, or merely useful for providing insight to subject experts, must be rigorously evaluated. Quantitative and qualitative models may use a wide variety of different mathematical approaches and representations, requiring a wide variety of technical expertise, e.g., agent-based simulation, dynamic systems models, social network models, regression models, hidden Markov models, etc. Regardless of the specific mathematical underpinnings of a model, the following questions should be answered as a starting point for the evaluation of any model.

- What specific problem is the modeling effort trying to solve?
- What simplifying assumptions does the model make?
- What information, quantitative and/or qualitative, will be used?
- Are the model's inputs and parameters knowable quantities?
- How much do perturbations of the inputs and/or parameters affect the model's output predictions?
- How will the modeling add value to the cumulative understanding of the problem?

It might seem that answering these questions should suffice for the evaluation of a model. However, there are more things to consider that have subtle nuances depending the use of the model, for prediction or insight. The following table provides the additional criteria.

Models for Prediction	**Models for Insight**
Have the predictions of the model been compared to real-world data?	Have model insights been corroborated with other models and experts? Do experts (other than the models developers) think the model helps them?
Were the test (validation) data independent of any data used to develop the model?	Has the model been used prospectively on new data or problems, or only on existing datasets where real-world outcomes are already known by the models developers?
When prediction accuracy is reported, were all the model's predictions considered (both good and bad ones), or does the investigator only highlight successful predictions?	Do experts rely on the model to commit to new operational decisions, or do they only rationalize real-world results against the model afterwards with 20/20 hindsight?
Are the test data (which were almost certainly derived from an existing, retrospective dataset) acceptably representative of the future events we want to predict?	Is the model capable of suggesting a wide range of real-world outcomes including counterintuitive surprises, or does the model make assumptions that build in a reassuring preconceived bias toward an expected outcome?
If the output prediction is subject to uncertainty, does the model report this uncertainty?	If the model (e.g., a strategic gaming exercise) might yield a different outcome if it is run more than once, has it been run more than once to see the range of possible outcomes?
Are the outputs actionable - does the prediction suggest meaningful ways of reducing the probability of the event, or ameliorating its consequences?	Does the model provide non-obvious insights that help experts make better operational decisions, or do experts do just as well without the model?

It is surprisingly difficult to evaluate models, more specifically quantitative models, reliably. The problem is that at any given time, it is only possible to evaluate a model against events that have happened. It is surprisingly easy to artifactually "predict" events after they happen. The various items listed above help guard against different ways of fooling oneself.

For example, it is common to split datasets into independent "training" and "testing" data for parameterizing then evaluating a model. Insidiously, this evaluation protocol only works *once*. If an investigator observes a result on the test dataset and responds by doing more work and making "improvements" to the model, now the test dataset has been used as additional training data, and the evaluation protocol is no longer reliable. One way to help guard against this is sensitivity analysis. For example, if a model's parameters are overtrained (if they have inadvertently "memorized" the test data in some way) sensitivity analysis may reveal that the model's predictions seem inordinately sensitive to an unreasonably precise choice of model parameters.

Another way that predictions are artifactually evaluated is when many predictions are made, and only the successful ones are highlighted. A model that has only 1% accuracy may still make a successful prediction if 100 predictions are made, perhaps in one paper, or perhaps in one lucky paper in a field of 100 papers. (This concept is related to the topic of false alarm rates discussed in Section 5). It is easy to focus retrospectively on the successful prediction, and lose sight of the actual poor predictive accuracy of the model. Rigorous effort must be made to define what specific predictions have been made *before* evaluating their overall success or failure.

From a research program's standpoint, it is essential to set the expectation that investigators evaluate their models. Because an investigator may take all reasonable steps and still be honestly fooled (for models more complex than simply predictable physical laws), it is also essential to demand that investigators freely share their datasets to enable the performance of different methodologies to be rigorously compared by independent investigators. Programmatically, these principles are the only way that a field gains traction and moves forward toward building better models.

4.2 Data

Data in and of itself is never useful, data must be coupled with a model or models in order for it to provide actionable information. This coupling occurs on many levels: determining what sort of data should be acquired (e.g. family trees, website postings, phone logs); focusing resources on select subgroups of data (e.g. only some website postings); and interpreting the data (what level of risk is associated with a particular observable or set of observables). The relationship between data and modeling is addressed in this section.

Section 2 argued for adding motive to the WMD-T event assessment. One advantage of shifting from a focus on intent to a focus on motivation is that data on motivation may be more widely, openly, and eagerly distributed than information in intent. Covert acquisition is not required, but raw data alone is of no use, models, analysis and interpretation are required. Figure 2 provides useful and complementary shaping concepts for collecting and categorizing information for evaluating the potential for terrorist action.

Oddly, counterterrorism is an area where there is both a paucity of data and an embarrassing richness of data. There is a paucity in the sense that there is almost no available data on major terrorist attacks. Evaluation based on models matching such small data sets is almost silly. On the other hand, there is a tremendous amount of data that might contain extremely useful information, but which dominantly contains less useful information. For example, video cameras recorded Mohammed Atta and the London bombers before their attacks, but such cameras record many millions of images every day. The clutter surrounding the useful data is so large that it is impossible to separate out the meaningful information even when it is clearly available.

Any algorithm that allows one to eliminate useless data would be extraordinarily important since it would reduce the clutter that surrounds useful data.

In data acquisition, different strategies are required depending on ones models and goals. Three obvious levels of threat monitoring and countering are:

- regions or groups considered likely to be the source of future terrorist activity, (e.g. failed states, groups with serious ethnic grievances);
- actual particular organized groups with a known or suspected willingness to commit terrorist acts, and
- individuals who have characteristics that are considered to create a propensity for terrorist attacks, (e.g.,membership in organized groups with known or suspected willingness to commit terrorist acts.)

For data at the level of regions and groups, public information, polling services (see Section 2.3), and covert intelligence may all play a vital role in acquiring information. For information on actual groups, polling data may be available for political wings, but is unlikely to be available for the militant wings. However, militant wings often make public statements that do provide substantial information about motivations and goals. Polling data on individuals is meaningless. Open data on individuals is available in public records. Some individuals, like Osama bin Laden, elect to make public statements, but for many individuals public statements are unavailable. Thus, data mining of public information and covert data acquisition may be the most important tools for collecting data on individuals.

Models developed for questions about terrorism may be based on many factors. For example, some models suggest that the following factors play a role in determining whether or not an individual person will commit a terrorist attack: attitude toward authority, religion, degree of devoutness, gender, age, relationship with parents, familial structure, marital status, history of aggression, employment history, education, drug use, group membership, patriotism, reading habits, favorite websites, and hobbies. Some of this information is readily available in the public record, and other parts, such as the persons emotional relationship with his parents, may be unknowable without substantial direct questioning of individuals. The NNDB [5] data base [12] is an open website that posts much of this information about people who have committed terrorist acts.

As stated earlier, model results depend on the assumptions underlying the model. One of the largest assumptions is the universality of the data set. Models gain validity by being tested against data that has never been seen. If the data set is not universal, but contains particular elements with different features, models will become very confused. Particular terrorist groups have individual features that are particular to their goals and tactics. It is very difficult to be certain that all of the data really belongs to the same data set. It is not clear that information on the behavior of the IRA is universally applicable to al Qaida, though some elements of tactics map very well from one to the other.

Similarly, time is an important variable. Natural evolutions with time, such as the advent of cell phone and internet technologies have radically changed communications methods for terrorist groups. In addition, terrorist

[5]The actual validity of of the NNDB is unknown to JASON and should be evaluated if these data are to be used in modeling efforts.

groups actively respond to countermeasure, as clearly demonstrated in the IED problem in Iraq. Thus, using an old data set to project future behavior may be produce erroneous results.

Data can never make recommendations. Carefully evaluated models are required to convert data into recommendations. These recommendations are best if the assumptions underlying them are made clear, and if the connections between the actual raw data and the recommendation are clearly laid out with explicit statements of uncertainties. Data acquisition and modeling will produce the most useful recommendations if they are focused on the specific questions being asked by policymakers. "Is building roads more important than establishing dispute resolution mechanisms in villages?" This is a much more tractable question than"What can we do to win in Afghanistan?" "Does offering rewards to the families of people who renounce terrorism decrease the rate at which people return to terrorist activities after their release from custody?" This is also much more tractable than "How can we end terrorism?"

4.3 Data Sources

The availability and applicability of data sources for the WMD-T assessment is of concern. Even for straightforward data-driven question as a plot of event frequency versus magnitude given in Section 3.3, it was surprising hard to obtain primary datasets and reproduce published results like the power-law fit shown in Figure 5.

Several datasets of terrorist events have been collected. Only one, the National Counterterrorism Center's Worldwide Incidents Tracking System [8], appears to be openly available. This database only covers 2004-present.

Three datasets are available through a registration (subscription) request process. These include: the Global Terrorism Database [10] from START[6]; the ITERATE database [11] at Harvard University, and the RAND Corporation Database of Worldwide Terrorism Incidents [9]. Another widely used database, the Terrorism Knowledgebase at the Memorial Institute for the Prevention of Terrorism in Oklahoma City, was defunded and closed in 2008 [6]. There are also at least three sources of "chronological narratives" that provide more detail per event, but not in a parseable form: the State Department's annual reports, which track non-US events; the FBI's annual reports, which collected domestic events only, and appear to have ceased in 2005; and the National Counterterrorism Center's annual reports, which started in 2007 and appear to cover all worldwide terror events.

Dataset availability, enabling reproduction and extension of published research, is a significant concern in many other fields, particularly in other highly empirical (data-rich/theory-poor) fields like the life sciences. A 2003 National Academy of Sciences report [5] affirmed the general scientific community's expectation in the life sciences, if not all science, that

> "... the general principle [is] that the publication of scientific information is intended to move science forward... An author's obligation is not only to release data and materials to enable others to verify or replicate published findings (as journals already implicitly or explicitly require) but also to provide them in a form on which other scientists can build with further research."

This leads us to end this section with the following finding and recommendation.

[6] A national consortium for the Study of Terrorism And Responses to Terrorism, a Department of Homeland Security center of excellence at the University of Maryland.

Finding: The availability of high-quality published datasets in social sciences related to WMD-T appears to fall short of the open standards expressed in this 2003 National Academy report [5].

Recommendation: Researchers should make their primary datasets openly available to other researchers for reproduction and extension of results.

5 THE FALSE POSITIVE PROBLEM

Any problem involving prediction of rare outcomes, including medical diagnoses and disaster warnings, needs to pay careful consideration to *false positive* prediction rates, i.e., *false discovery rates.* False discoveries lead to false alarms. A model that overpredicts may be worse than useless, distracting planners and forcing expensive responses to false alarms. The same should be true for the assessment of WMD-T events. In the terrorism literature there are may retrospective claims of successful "predictions" that do not take into account how many other equally valid predictions had been made, including claims about the events of 9/11.

Finding: Insufficient attention has been paid to defining false discovery rate for any of the components of rare event assessment.

Recommendation: Require the evaluation of false positive prediction rates for the putative rare event left-of-boom detection systems and of the impact of these rates in a real-world deployment of the systems.

5.1 What are False Positives?

The working hypothesis underlying the SMA effort is that one can formulate a technical strategy which uses data from both public and other sources and detects a terrorist group attempt to create a large scale incident. Needless to say, all this data will be buried in a very much larger data stream and the problem at hand becomes very much a question of finding methods that

can deal with rejecting the vast amount of background in favor of the signal. Basic detection theory characterizes the functioning of this class of detection strategy by means of the $precision - recall$ curve which plots $precision$ or one minus the false positive rate, (*1*-(false discovery rate)), versus the detection probability while some feature of the detector, perhaps the threshold at which a detection is said to have occurred, is varied.

Figure 7 gives the definition of $precision$ and $recall$, along with the curve. This curve makes it clear that detection capability, as determined by the probability that a true signal will be detected, is always balanced by the probability that a false alarm will be declared. Without determining the acceptable value of the false discovery rate, talking about detectability is meaningless. One could declare all input data to be a detection and assure that $precision = 1$, but without any useful discrimination capability. This approach would be clearly useless, so it is never done. Instead, one gets the kind of $precision - recall$ curve shown in Figure 7, with $precision < 1$ at $recall = 0$.

For a system that will be monitoring various signals in continuous time, the correct measure becomes the false discovery rate. JASON found almost no discussion of what might be an acceptable value of this rate for the WMD-T scenario of trying to use both public domain and intelligence data to determine when a group has reached the stage where they both intend to commit a major act and have established a plan to procure the necessary materials. Since any discussion of detectability cannot even begin without some notion of what is an acceptable false discovery rate, we find this to be a major omission in the SMA framework.

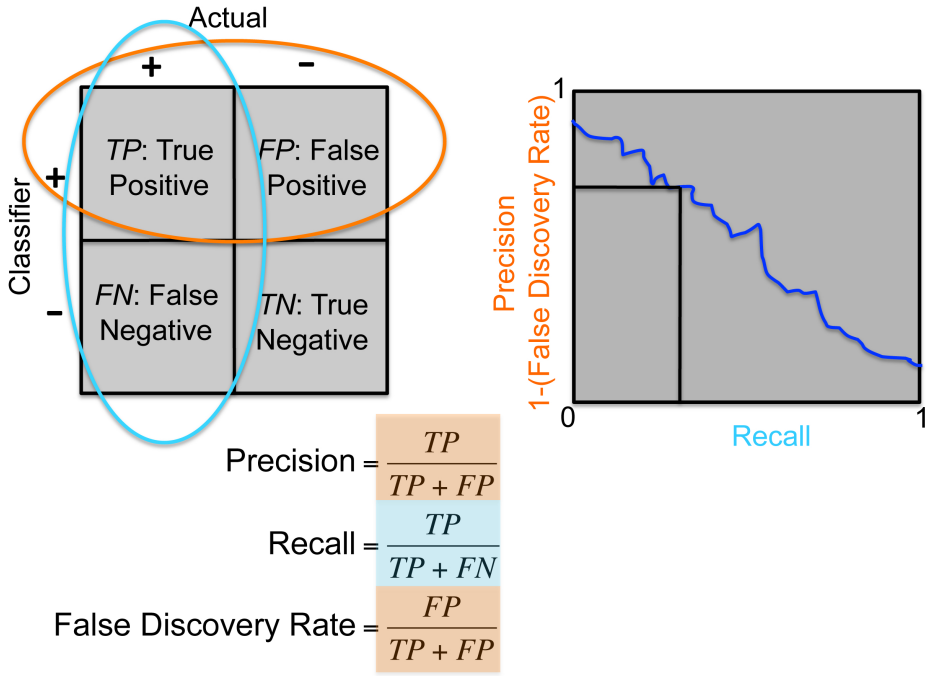


Figure 7: Precision-Recall Definitions and Curve

An interesting anecdote from earthquake detection illuminates the basic point here. During the cultural revolution period, it became Maoist doctrine that earthquake prediction could obviously be accomplished through the "unfailing efforts of the broad masses of people" ([27], [28].) During this era the Chinese government made a successful prediction of a 7.5 magnitude quake in the city of Haicheng, apparently based both on physical precursors and on traditions regarding anomalous animal behavior prior to quakes (according to some Japanese, catfish are more capable than seismologists for predicting earthquakes). But, there was no prediction of the 1976 quake that killed one quarter of a million people in Tsanghan. More to the point here, there were over thirty cases of widely publicized earthquake warnings in the late nineties, none of which proved accurate. The cost of these alarms, in terms of evacuations and business disruptions were so severe, that the Chinese government issued new regulations requiring "a high standard of

scientific reasoning" for all predictions and furthermore introduced penalties for inaccurate warnings.

As the above example illustrates, the cost of a false alarm depends on exactly what will be done whenever an alert is triggered by the putative rare-event "left-of-boom" detection system. Obviously, responses can range from assigning a specific cohort of people to investigate the case more fully, all the way up to a public disclosure and evacuation plans. It might be reasonable for a system to have, just to pick a number, one false alarm per month if this leads to activation of more serious (i.e. more expensive) tools and these tools can decide if this was a true detection or not. But, this high order process would need to be specified in terms of its $precision - recall$ curve.

Once one has established a reasonable bound on the allowable false discovery rate, one must attempt to maximize *precision* without violating this constraint. This obviously places the onus onto developing a set of event precursor signals which relatively uniquely point to a large, perhaps WMD-T based, event in the offing. Exactly how hard or easy this might be to accomplish is not at all clear, but the rarer the event, the more difficult this will be.

There is a dilemma here in the terms of thinking about the connection between rare events and more common terrorist events of much smaller size and impact. As mentioned earlier, one possibility is that rare events are just larger versions of typically-sized cases, the "largeness" coming from a random assortment of factors that happen to make the consequences large. The 9/11 attack might be construed as being of this form, as the size of the casualty count was due to unforeseen cascading of factors (the destruction of the building could not have been expected) and was ultimately limited by the surprisingly efficient evacuation. An earthquake analogy here would

be that large magnitude quakes arise from the same types of initial events as is the case for typical events, but that there is a random cascading of energy release due to unforeseeable details which is stopped by yet another unforeseeable combination of factors. Evidence for this type of connection might arise from having large events fall on the same size-frequency plot as the small and medium sized ones, as discussed in Section 3.2. If this is true, one can indeed use the more plentiful small events to extrapolate to what might be seen for large events. But, given that they involve similar underlying processes, there will be an issue discriminating on the basis of event size based on early indicators. This appears to be one of the most severe problems in using "pre-shocks" to predict earthquakes. This is because it is only apparent after the fact which pre-shocks end in a whimper and which lead to a bang. Prediction approaches which just raise the alert level based on the occurrence of any sufficiently large shocks (the so-called automatic alarm strategy) are plagued by unacceptable false alarm levels.

5.2 Strategies for WMD-T False Alarm Discovery

One strategy for having a meaningful detectability within the false alarm constraint is to assume (or hope) that rare events really are of a different breed. For example, one might postulate that WMD-T based events, especially including nuclear explosions, might just be completely different in terms of the level of planning, the needed amount of information and material gathering, as compared to run-of-the-mill TNT based attacks and therefore one would be unlikely to confuse the respective indicators. However, we have practically no data of real events upon which to base our detector strategy. This means that we have to rely almost exclusively on theory-based scenarios

and on lessons learned from artificial exercises such as the Limited Objective Experiments (LOEs). LOEs, a concept introduced in 2001, are multiplayer experiments designed to exploit and study information sharing and collaboration ([35], [37]). JASON found WMD-T LOE concept and related games do not appear to be optimally designed for the purpose of giving the necessary insight into the false positive problem. (A specific WMD-T SMA LOE will be discussed in detail in Section 6.1).

Given the lack of real-world examples of the type of rare events under consideration here and given the hope that they are actually distinguishable from the large background, red-team based games play an essential role in figuring out what may be possible. Games solve the problem of not having the ability to reconstruct all the events that occurred prior to a specific attack in the past (and the even more serious problem of wanting to understand possibilities for which we have no prior data) by erecting a managed time line in which events are arranged to occur in a manner that allows for capture of all possibly relevant information available prior to "boom". Given that these games last for a short period of time, the rate of event occurrence within the game are wildly unreasonable. For example, during the SMA LOE no fewer than three events are supposed to either occur or be in progress. This converts the detection problem into an "answer in the back of the book" exercise because all participants know that there is a real event buried in the data as opposed to the real-world case where there will almost always not be a real event embedded in the ingested data at any given point in time.

The other difficulty with the LOE style exercises from the point of view of detection theory is that the game is built upon the pre-conceived notion that the road to detection is via collaboration. Essentially, enough information which taken together as a clear attack indicator is distributed piecemeal

to the participants. If they can figure out how to collaborate so as to gather more of the puzzle, they can detect the planned event. The game is a scavenger hunt for the right clues. This idea seems to be based on impressions that the real problem of predicting 9/11 was that different agencies had different parts of the complete picture and if these pieces had been combined (i.e., had only the administrative barriers been absent), the attack would have been detected and prevented. This might be true but just as likely might not be.

As we have already discussed, it is always possible to pick out the important clues retrospectively, just as it is always possible to pick out pre-shocks of a large quake. Whether they are unique enough to find them out of the background needs to be tested in a game that has realistic ratio of true clues to false ones and where the correct answer most of the time is that there is no imminent, credible threat. Perhaps the fact that the infrastructure developed for the LOE will be maintained for continuing use can be utilized to run the game many times over, with events inserted only sparingly, and with a more dynamic red team involvement. This would allow one to move away from a pre-determined scenario of what needs to be detected. Having a dynamic red team would help address the obvious concern that terrorist planners would arrange their actions so as the generate as few observable signals as possible. In some cases, they might even produce false signals on purpose to try to create false positives to cause us to raise our detection threshold.

One interesting direction for future work might be to embed rare event detection games into virtual worlds such as Second Life [33]. In fact, there have already been examples of terrorism in these games. For example, the Second Life Liberation Army has a list of grievances against corporations who have populated this digital universe and have staged a series of bombing

attacks [38]. From the rare-events perspective, these multiplayer role-playing communities are more natural settings for social experiments than are the one-shot constructions used for the LOEs. One might therefore imagine working together with the companies that manage these games, Linden Labs [34] in the case of Second Life, to develop a terrorism detection capability, either using actual in-situ terrorist groups or via inserting red teams designed for this purpose. The fact that these multiplayer worlds would have millions of participants and are ongoing enterprises might allow for an evaluation of detection algorithms under much more sociologically realistic conditions.

There is one point at which the direct analogy with detecting physical signals breaks down making the WMD-T problem even more difficult. Absent theories of a malevolent deity, the earthquake is not aware of our "left-of-boom" detection attempts and does not work to oppose them. Yet, this is exactly what might be expected of a terrorist group if the methodology being used to detect planned incidents becomes known to it. To take a simple example, Mohammed Atta was quite visible on the video camera of the Maine airport when he entered security, but there was no way to attach significance to this image in the midst of an enormous number of a priori identical ones. Imagine that it was known to, or even just assumed by, the terrorist group that a highly accurate biometric system was able to detect everyone on a watch list that passed through an airport and that such an event would cause an airport shutdown, preventing any airplane hijacking. Even assume that we were willing to live with the enormous numbers of false alarms such a system would create. Any terrorist would then be strongly motivated to engage in simple acts of deception such as facial disguise that would defeat the detection strategy. In that case, and if there is enough time to plan counter-moves, game theory predicts that the detection problem becomes

essentially unsolvable. This is discussed in detail in Section 6.3.

Charles Richter, probably the most well-known American seismologist, is alleged to have said "No one but fools and charlatans try to predict earthquakes". Much of the work on rare terrorist events seems to take for granted that "the truth is out there" and we can discover it in a sufficiently timely fashion with the right mixture of motivational assessment, social network analysis, capability measures etc. Perhaps this is the only approach which is politically defendable, but it may be the case that achieving a high probability of detection in the presence of an immense background and in the presence of informed terrorist attempts at signal concealment is simply not attainable without having such a high false positive rate as to make normal life impossible. The lack of attention to any quantitative measures of this problem makes it impossible at present to assess the likelihood of this pessimistic possibility.

5.3 Lessons Learned from Near Earth Objects Community

Another community worth drawing some useful analogies from, in the context of WMD-T detection and information sharing is the Near Earth Object (NOE) Community. The NEO community includes two groups of people, astronomers who study the population of natural objects in space close to the earth, and concerned citizens who study the consequences of collisions of such objects with the earth. The main data-base of the NEO community is the International Astronomical Union Minor Planet Center (MPC) at Harvard University. The MPC makes public announcements when new NEO are discovered and when observations of NEO are made which allow

the probabilities of impacts to be calculated. NEO impacts are like terrorist attacks in causing major catastrophes with low probability and minor catastrophes with high probability. NEO impacts differ from terrorist attacks in having probabilities that can be accurately calculated from observations of the NEO orbits. Another difference between impacts and attacks is the contrasting attitudes of the two communities of experts toward sharing of information. The experts on terrorist attacks, who mostly belong to national police or intelligence organizations, are traditionally secretive and maintain data-bases which are compartmented and not widely shared. The NEO community has a long history of cooperative sharing of information, starting in the year 1801 when the first asteroid was discovered.

Especially in Europe where modern astronomy began, skies are often cloudy, and the discoverer of a new object depends on foreign colleagues to keep the object under continuous observation and establish its orbit. An important event in the history of astronomy was the successful laying of the Atlantic telegraph cable in 1866. As soon as the cable was working, the Astronomer Royal in London and the director of the Harvard College Observatory in Massachusetts began to use it to send regular telegrams reporting new discoveries on both sides of the ocean. In 1882 an official institution for the international sharing of information, the Central Bureau for Astronomical Telegrams (CBAT), was established with headquarters at the Observatory in Kiel, Germany. The CBAT still exists, although it now distributes announcements of new discoveries over the internet instead of by telegram, and its headquarters is now at the Harvard College Observatory. The MPC and the CBAT work side by side, the MPC collecting information about small objects in the Solar System, and the CBAT collecting information about the rest of the universe. The tradition of openness and prompt international

cooperation, epitomized by the CBAT, has been maintained by the NEO community. Nothing comparable exists in the community of experts on terrorism.

A typical NEO is discovered with a poorly determined orbit which gives it a low probability of impact. As time goes on, new observations are made, the orbit becomes more precisely known, and the probability of impact slowly increases. Then rather suddenly, the orbit becomes so precise that the probability of impact becomes either one or zero. In the vast majority of cases, the probability becomes zero and the object misses the earth. For the typical NEO, the probability of impact rises slowly to a small maximum and then falls abruptly to zero.

Both for impacts and for attacks, any attempt to give advance warning of catastrophic events to public authorities is bedeviled by the problem of false alarms. After a few episodes of disproportionate response to alarms that turn out to be false, public authorities cease to respond seriously to alarms that may turn out to be true. To deal with the problem of false alarms, the MPC invented a scale called the Torino scale which is supposed to describe the magnitude of the risk of impact posed by each NEO. The risk is measured by a single number which combines the size and the probability of the impact. Each announcement of discovery or observation of a NEO is accompanied by a number on the Torino scale which begins low and rises slowly, so that there is no sudden alarm. The Torino scale was introduced in 1999 and replaced by another more accurate scale called the Palermo scale in 2008. The Palermo scale is still a single number, but it takes into account the date of a future impact as well as its size and probability. As the date comes closer, the risk becomes more urgent, and the number on the Palermo scale becomes larger.

Although the Palermo scale gives increased emphasis to near-future impacts, it remains true that the very large far-future impacts usually dominate the overall estimates of risk. At the present moment the dominant risk on the Palermo scale is associated with asteroid 1950 DA, a kilometer-diameter object which is calculated to impact Earth in the year 2880 with probability 0.003. Fortunately, we have plenty of time to measure the orbit more accurately and to see the probability of impact rise to one or fall to zero. In case the probability rises to one, there will probably still be time to deflect the object with a small steady thruster powered by solar energy, to make sure that the object will miss the Earth. The power required to do the job is only a few kilowatts if the thrust is maintained for 400 years.

Unfortunately, the study of terrorist attacks is not an exact science like celestial mechanics. There is no Palermo scale which allows us to identify a dominant risk of a catastrophic terrorist attack 800 years in the future. But the study of NEO impacts may still teach us a useful lesson for dealing with the problem of terrorism. The lesson is to think of the problem with a time-horizon of centuries rather than years. The dominant risks of terrorist attack may come from large-scale and devastating attacks far in the future, and we have a better chance to avoid such attacks by long-range transformations of society than by short-range tactical counter-measures.

6 GAMES AND GAME THEORY

The previous section began a discussion of the need for games and expertise for the study of detection rates of rare events. We take a deeper look at these issues here. In addition, we develop a game theoretic framework to model defenses against WMD-T events.

6.1 Limited Operational Exercise

JASON reviews an upcoming limited operational exercise (LOE) sponsored by SMA. This is an example of the use of a multiplayer game to investigate aspects of the rare events detection process. This exercise involves individuals from a variety of government agencies who will play two hours daily for a month. The idea behind the game is to introduce information regarding terrorist attacks into otherwise normal data streams and see to what extent information-sharing among the participants can be used to put together pieces of the "puzzle" and decipher the clues.

There are certain aspects of this exercise which are quite interesting and well-thought-out. The idea of embedding the clues into a web portal system which mimics the normal work environments of different analysts is a major improvement over the more artificial interfaces typically encountered during games. The use of the real internet "swamp" as the playing ground allows for a realistic "clutter" against which the signal must be detected. Also, the game will last for one month, as distinct from the more typical ones which just take a few days, allowing for a somewhat more realistic timeline for scenarios to unfold.

What will be learned from the LOE and equally important, what will not be learned? The LOE designers have postulated that left-of-boom detection is similar to a puzzle in which the main challenge is for individuals to find pieces and then identify and communicate with other players so as to collaboratively put the pieces together to solve the puzzle. In other words, the hard part of the problem will be to identify who among the other players can help an individual notice and then interpret information that they have stumbled across, which by itself would not be sufficient to understand what is going on. What can be learned from this relates to institutional boundaries for information sharing, the possible types of incentives that could be used to overcome these barriers, and more generally the type of social networks that would "self-organized" in response to the posed challenge. As a social science experiment it seems well-designed and well-implemented. Also, as a response to the folk wisdom that the lack of inter-agency information sharing was a major contributor to the 9/11 intelligence failure, it seems useful. There are many important issues that will not be addressed by this LOE. These include:

- There is a difference between information-sharing and collaboration. The latter involves much more than looking someone up in a directory and sending them a request for information or a request to help interpret some data. Especially when individuals start out with different areas of technical expertise, it can take many exchanges over long periods of time to reach a true collaboration in which truly interdisciplinary issues can be addressed. This has been the experience of all scientists who work at disciplinary boundaries. If the rare-events detection process actually requires interdisciplinary collaboration (a reasonable

hypothesis), the LOE will fall short in providing information regarding how to create such collaborations.

- As mentioned in the previous section, the rate at which events occur in the LOE belies the notion that it is a mechanism by which we can study rare events. It will not lead to any quantifiable estimates of detection probability and false positive rates.

- There is no active red team involvement. Red teams are necessary in order to capture the feedback dynamics by which terrorist groups modify their plans and methods (and hence the detectable signals they generate) in response to their perception of the blue team's counter-terrorist strategy. This feedback makes the detection problem much harder than might be the case for physical examples of rare events (for example, increased surveillance could be counter-detected and could lead to measures being taken to disguise planned actions) and also may have surprising consequences for very long term prediction (i.e. prediction based on an analysis of motivation and capability, in the absence of any specific actions that might be detected); this will be discussed from the perspective of game theory in a later section.

Our fundamental recommendation is that the SMA should develop additional exercises that focus on some of the issues not addressed by this LOE and should carefully consider how each game will lead to increased insight. The fact that the enabling technology for the LOE will be maintained should provide invaluable in this regard. This discussion leads to the following finding and recommendation.

Finding: Focus seems to be on only a limited number of the compelling issues.

Recommendation: Make clear statements about what is to be learned from the exercises and expand the potential set of issues to study.

6.2 Red Teams

Here we discuss more specifically some of the issues that arise upon consideration of red teams (RTs) and games. It is clear to us that games in general and more specifically strategic games involving RTs should be utilized to help with WND-T rare events assessment. The basic idea is to form teams (typically 6-10 people) who can accurately represent the culture, motivations, and capabilities of a terrorist planning group (TPG). If this can be accomplished, the actions of the RTs can assist in following aspects of the problem.

- RTs can help discern TPG priorities among various targets as objects of large scale terrorist attacks. This type of assessment does not involve detecting actions that arise as part of an actual planned attack, but instead tries to use expert information regarding terrorist motivation to prioritize various targets. As will be seen in the next subsection, this type of information is extremely important when planning our defensive strategy.

- RTs can help determine potentially observable physical signals of TPG-directed pre-event activities. This should be done in the context of

background clutter, which may make observable signals difficult to notice.This is certainly a necessary part of any detection scheme.

- Unlike the case for physical events such as earthquakes, TPG can react to defensive activities (e.g. changing local defenses of certain attractive targets or better monitoring by us of some TPG activity) at various stages in their pre-event attack preparations. Red teaming is basically the only viable approach for understanding these feedbacks.

Forming optimally useful and reliable RTs is crucial to these goals. Important questions remain to be answered. How big should a RT be? Should it be about the size of the TPG or larger to encompass more possible viewpoints? How many independent RTs should be formed to choose and plan a future large impact terrorist event? Differences among RTs in making event priority lists would indicate the uncertainties we should assume in predicting TPG intentions and event choices. What are the personnel and budgetary constraints on this number?

Perhaps the most crucial question is what criteria should be used in choosing RT members? Large differences may exist between the TPG to be imitated by the RT and the pool of experiences from which the RT members are typically chosen, (e.g., cultures, religions, possible fanaticism, upbringing, ethnicity, etc.). Special efforts should be taken to minimize these differences by including experts in the relevant foreign culture(s) in the RTs. However, it was cogently argued in a JASON briefing that it is even more important that the RTs reproduce a narrower "professional culture" of the TPG they want to imitate (e.g., knowledge, experience, skills, education, training, and contacts).

The optimal composition of the RT may also depend sensitively on which special stage in the long life of the TPG one is trying to mimic. The TPG may initially form with the objective of simply deciding the nature and future timing of a large scale terrorist event. After this stage, the original TPG may then invite others to join with special expertise, (e.g., experience in explosives, or biological weapons, or radiological ones, others who know more about the defense of special targets in the US). All of this planning may take many months, perhaps even a few years. At that point, some of the earliest TPG members may no longer be part of group, at least not as active decision makers. How should the RT best accommodate these changes? How best to choose and perhaps change RT composition in a group that probably meets together over a small time interval,very much shorter than that of the TPG group it tries to imitate.

One may be optimistic that future studies optimizing RTs would give more reliable answers to some of these questions than is available today. As we have argued, these could have very important consequences for some studies of the risk and or predictability of terrorist initiated rare events. This leads to the following finding and recommendations.

Finding: Crucial questions about size, composition, and duration of red teams remain open.

Recommendations:

1. Optimize the structure and use of red teams, both strategically and tactically.

2. Replication of exercises is essential to documenting the reliability of the results.

6.3 Game Theory

One approach to the rare event problem is to try to predict based on terrorist culture and motivation exactly which targets would be most tempting as a venue for a large-scale attack. This type of prediction would require bringing in relevant social science experts and working on a better understanding of motive and how different target choices might be better or worse choices. We would then allocate our defensive resources to try to counter those attacks, obviously concentrating more resources on more likely high-value targets. But, it is vital to recognize that the problem is not just "feed-forward". The terrorist recognize what targets we are defending and change their attack priorities accordingly – it would be better to have an assured success with a second or third choice target than to have a negligibly small chance of hitting the big one. If there is enough time, the two sides will reach an equilibrium of defensive choices versus target list. This can be analyzed by the methods of game theory. The interesting result that will emerge below is that the predictability leverage that exists is solely due to *differences* in perceived target value and that therefore is an area towards which social science research should be directed.

6.3.1 Undefended target values and zero sum

Explicitly or implicitly the terrorist assigns values V_i, $i = 1, \ldots, N$ to the targets, and also evaluates probabilities p_{ui} that an attack will be successful

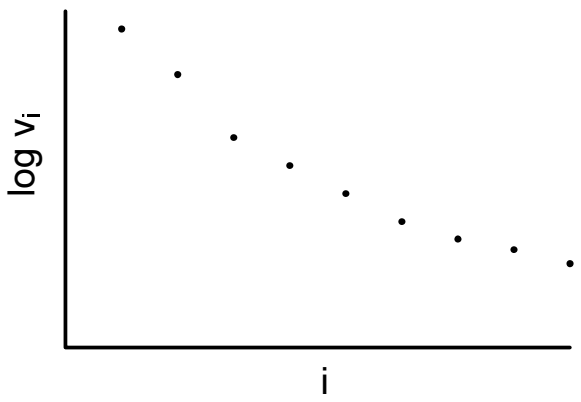


Figure 8: Potential terrorist targets are ordered by "net" value v_i and plotted on a logarithmic scale. Here we assume that we and the terrorist agree on the value of each target (zero sum assumption).

in the case of the target being *undefended*. Taking these probabilities into account, the expected value of each undefended target to the terrorist is

$$v_i \equiv p_{ui} V_i. \tag{6-1}$$

Of course, real terrorists may not be motivated exactly by linear expectation value. So, in what follows, we'll only use the values v_i, not V_i, with the understanding that the v_i's are the "net" values to the terrorist of undefended targets, after taking into account the probability of success.

In this section we explore the case where the game is zero sum. That means that we and the terrorist are in agreement on the values v_i. Any gain for the terrorist is considered to be an identical loss to us. In Section 6.3.2 we'll relax this assumption, with interesting consequences; but it is useful to understand the zero-sum case first.

Figure 8 shows a notional plot of the values v_i of targets, plotted on a logarithmic scale and ordered from most to least valuable.

Model for Defense with Fixed Total Resources: As a toy model, suppose that we have a fixed quantity D of total resources for defense, and that it can be allocated to individual targets in amounts d_i, with

$$\sum_i d_i = D. \tag{6-2}$$

We want to model target defense in a way that captures, if crudely, the ideas of (i) diminishing returns on a given target as we (over-) defend it, and, not unrelated, (ii) layered defense, where resources are allocated for defenses that come into play only after other defenses have been breached. One way of doing this is to model defense as decreasing the effective (expectation or net) value of a target to the terrorist exponentially with the amount of defense allocated to it. That is,

$$v_{ei} = v_i e^{-d_i}. \tag{6-3}$$

This model could, of course, be improved by putting target-dependent constants also into the exponential. However, Equation (6-3) has a particularly easy graphical solution that will illustrate some generic features.

Nash Equilibrium Solution In general, two-person zero-sum games have a Nash equilibrium that is optimal for both players [51]. That is, neither player can do better by departing from the Nash equilibrium solution, so both players will rationally adopt it. And, if one player is less than rational, the other can only benefit.

Figure 9 shows the geometrical construction of the Nash equilibrium for the game that we have defined. It works conceptually as follows:

- Start by applying defense to the highest value target v_1, thus pushing down its value to the terrorist. Because we have plotted target values on

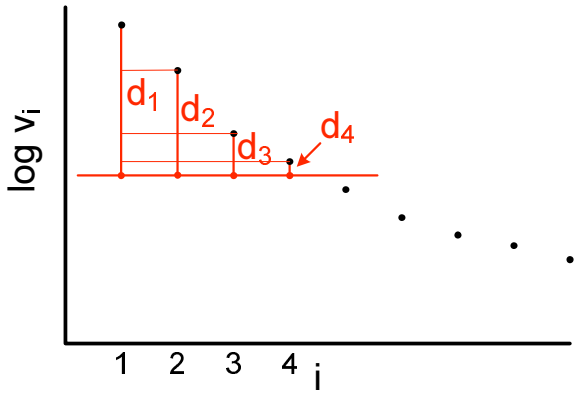


Figure 9: Construction of the Nash equilibrium defense for the targets shown in Figure 1. See text for details. Notice that at Nash equilibrium, multiple targets have *identical* attractiveness to the terrorist.

a logarithmic scale, Equation (6-3) implies that the amount of defense used is proportional to the distance that we push down the value on the graph. In this, and any subsequent step, stop when you have exhausted your supply of defense D.

- When v_1 has been pushed down to make $v_1 = v_2$, apply any remaining defense resources equally to push down v_1 and v_2.

- Ditto when $v_1 = v_2 = v_3$, apply remaining incremental resources to all three.

- And so on, for increasing v_i's with increasing i until all resources are exhausted. In Figure 9, this occurs when incremental defense is being applied to points 1–4, but before they have been brought down to the level of point 5.

Why is this the Nash equilibrium? Because any other strategy of allocating the d_i's would result in an end state with at least one higher (and highest) v_i, which, being the highest net value to the terrorist, would be attacked, producing greater loss to us than the Nash equilibrium strategy.

The terrorist's Nash strategy is to choose randomly among the defended targets (1–4 in the Figure) and attack that one. He can only do worse by attacking an undefended target, such as 5, so he has no reason to do so.

Predictability and Nash Equilibrium Before we applied any defenses, the toy model had a high degree of predictability: The terrorist will likely attack target 1, the highest value. If he attacks any other target, he achieves less.

After we apply defenses in the optimal Nash strategy, we know that the terrorist will attack one of the sites that we have optimally defended, but not which one. The fact that we don't know which defended site will be attacked is not an artifact of the toy model, but a very general feature of an optimized defense. The defense wouldn't be optimal if any predictability were left, because then we would be better off adding a bit of defense to the predicted target.

The important lesson is that when we and the terrorist agree on the value of targets, the more we optimize our defenses, the less we will be able to predict where an attack occurs. Incidentally, we also know that attacks on *undefended* targets are, paradoxically, *unlikely*, because their comparative values remain lower than any defended site.

6.3.2 Non-zero sum is different

We now look at the case where individual targets are valued differently by us and the terrorist, so that the game is not zero-sum: Depending on which target is successfully attacked, we may lose more or less by our evaluation than the terrorist gains by his.

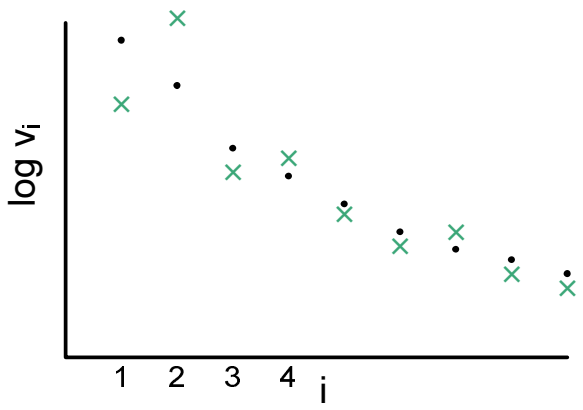


Figure 10: Notional target values that are different for us (black dots) than for the terrorist (green crosses). In this example, target 1 has the highest value for us, while target 2 has the highest value for the terrorist.

Value Gaps Figure 10 shows a notional non-zero sum case. The black dots (same as Figure 8) indicate our values for each target. The green crosses show the values to the terrorist, which can be greater (target 2) or less (target 1) than our values. As shown, before the application of defenses by us, the terrorist will attack target 2, because any other target has smaller value. Thus, to the extent that we can understand the terrorists values, we have perfect predictability.

Avoiding Unfavorable Degeneracy Because we know that target 2 will be attacked, the most efficient initial allocation of defense resources it exclusively to that target. We increase defenses until the green X is lowered to (almost) the level of the next highest green X, which is that of target 1. This is shown in Figure 11.

Notice what happens if we defend target 2 (red X) all the way to the (green X) value of target 1: The terrorist is now indifferent as to whether to attack target 1 or target 2. However, we are certainly *not* indifferent, since target 1's value to us (black dot) is much larger than target 2's defended value

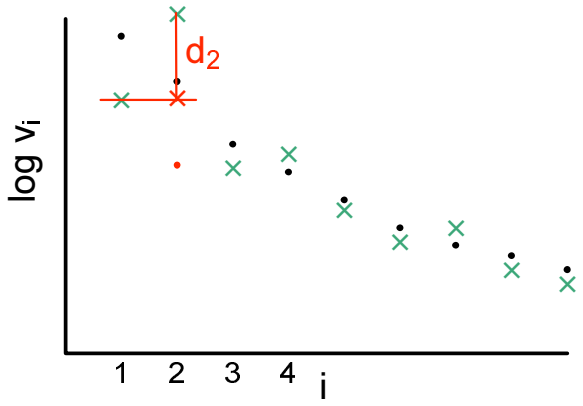


Figure 11: Initial allocation of defenses to target 2, which will predictably be attacked. If we defend until target 2 and target 1 become equally attractive to the terrorist, a discontinuity is introduced, to our detriment.

(red dot). It follows that, in this game, we must never create a degeneracy that can favor a target more highly valued by us. Instead, we should limit the defense of target 2 so as to leave it at least slightly more valuable to the terrorist. What "slightly" means will be determined by our level of uncertainty as to the terrorists values. We need to err on the side of safety.

Here the important lesson is to not over-defend targets more highly valued by the terrorist than by us when this may cause the terrorist to switch to a target more highly valued by us than by him.

The Rest of the Strategy Leaving the value to the terrorist of target 2 slightly higher than that of target 1, we now apply defenses equally to both, pushing them down as a unit (cf. Figure 8). We can continue until, as before, we hit a degeneracy that would cause the terrorist to switch to a target that we value more. Then, we back-off slightly and continue applying defenses now to all three targets equally.

In like manner, we proceed until our defense resource is exhausted. The result of this process is shown in Figure 12. Of the targets that we can afford

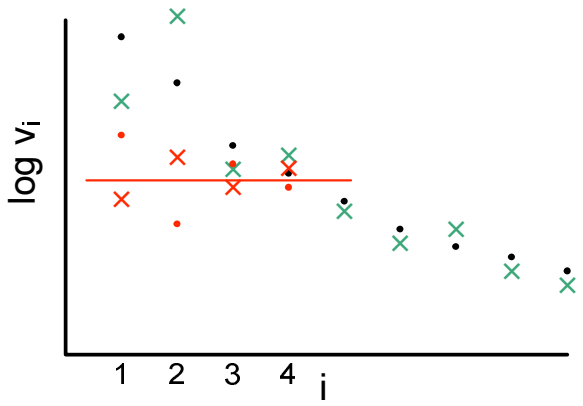


Figure 12: Optimal defense for the non-zero sum example. To the extent of our resources, we defend targets so as to order the terrorists preferences in the reverse of ours.

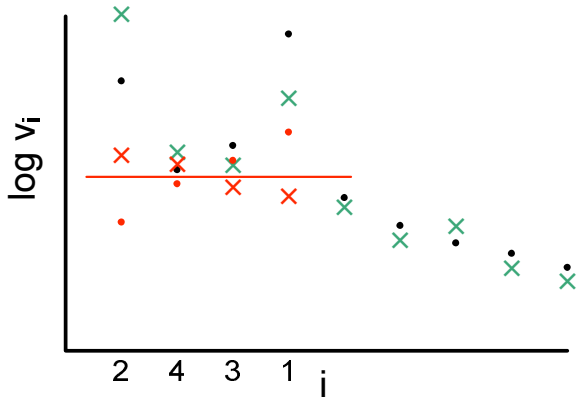


Figure 13: Same as Figure 5, but with the defended targets re-ordered by the size of the (signed) gap between our and terrorist values. The optimum defense orders the terrorist's values (red X's) into monotonically decreasing order, while ordering our values (red dots) into monotonically increasing order.

to defend, the one most likely to be attacked is, by construction, the one we care least about; and so-on for the other targets. This is made clearer if we re-order the four targets by the size of the signed gap between terrorist and our values, from most positive to most negative, as shown in Figure 13.

Engineered Predictability The key property of Figure 13 is that the red X's (terrorist values) are monotonically decreasing, while the red dots (our values) are monotonically increasing. Any allocation of defenses that makes this true is, in some sense, an optimal defense. What we would like to do is to make the magnitude of the slope of the X's as flat as we dare, so that the slope of the dots is as large as we can achieve. The terrorist would of course prefer things the other way around. He can affect our slope by trying to leave us uncertain as to his actual target values, so that we must err on the side of conservatism.

The following lessons are learned from this example.

1. It is important that we know as accurately as possible not only the values of targets, but also, specifically, the gaps between our and the terrorist's values.

2. The gaps are what allows us to manipulate our defenses so as to recover a degree of "engineered predictability" as to terrorist intentions. This contrasts the zero-sum case, where predictability vanishes for an optimal defense.

3. Exploiting the gaps gives not only predictability, but also decreases our potential losses by (in the best case) the size of the largest gap.

6.3.3 Secret selective defense

Up to now, we have made the assumption that the terrorist has perfect knowledge as to the amount of defense applied to each target. That is the safest assumption to make, because the defense of many targets is hard to keep secret; and we might be badly burned if we mistakenly assume secrecy in our planning.

Within the realm of the engineered predictability that is possible in non-zero sum games, however, secret additional defensive measures *on a small number of targets* may be useful. Secret defenses would be applied only to targets that, by the overt defense strategy outlined above, have a high probability of being attacked. The effect would be to push down the red dots in the Figures, without pushing down the red X's.

Here we see that by engineering predictability through the optimal application (and not over-application) of overt defenses, we also enable the use of secret defenses as an additional means of reducing losses. Notice that secrecy is much less useful in the zero-sum case because, absent predictability, we must apply secret defenses to a large number of equi-probable targets.

Secretly defended facilities are "honey pots", engineered to credibly attract attacks, because the terrorist recognizes that they are most valuable to him despite their lesser value to us (the gap); and to reduce losses because they are more strongly defended than the terrorist recognizes. One might think that in the zero-sum case, one could intentionally leave, say, one target "sticking up" as an attractive target, and then defend it secretly. The problem with this is that the terrorist can recognize the trap: that target *should* be better defended, so it probably *is* better defended by secret defenses. The

advantage of exploiting non-zero sum gaps is that the optimal overt defense engineers predictability in a way that is seen as rational by both sides.

This simple expose on game theory leads to the following finding and recommendation.

Finding: Game theory presents an interesting framework for resource allocation and planning.

Recommendation: Expertise should be tasked to understand value of targets from terrorist perspective.

7 CASE STUDY: BIOTERRORISM THREAT

We will end this report by reviewing what JASON has learned, from more than a decade of studies, about Bioterrorism. The challenge of predicting a major bioterrorism event provides a case study for quantitative approaches to rare events. Although there have been several small-scale events that might be classified as bioterrorism, there has not yet been a successful large-scale event. The two best known incidents, both involving the dispersal of anthrax spores, were the Aum Shinrikyo attack near Tokyo in June 1993 and the anthrax letters mailed in the Eastern U.S. in September through October 2001. The former was unsuccessful only because the terrorist group mistakenly employed a non-virulent vaccine strain (Keim *et al.*, 2001). The latter might not be regarded as a bioterrorism event, even though it caused seven deaths and incited widespread alarm, because it was carried out by a deranged U.S. government employee rather than a terrorist group. Nonetheless, can these and other small-scale events be used to predict a large-scale bioterrorist attack? What other quantifiable indices might be included in making such a prediction? How does one know that a hypothetical large-scale event derives from the same probability distribution as the observed small-scale events? Are there a sufficient number and diversity of small-scale events to provide a meaningful estimate of the probability distribution?

JASON has been studying the problem of bioterrorism for more than a decade and has conducted ten formal studies on this topic. These studies have considered aspects of intelligence gathering, capabilities analysis, surveillance, defense posture, response management, and forensic analysis. Throughout these studies risk assessment has been an implicit, but never explicit, component of the analysis. JASON's thinking has been guided by a

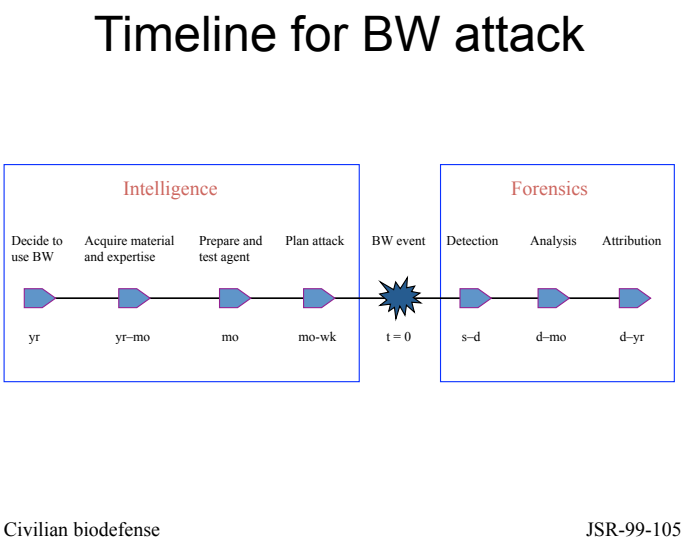


Figure 14: Biothreat Timeline

notional timeline for a bioterrorism event (Figure 14) that is analogous to the general timeline for WMD-T developed by the SMA program of the Rapid Technology Program Office (Figure 1). One notable difference is that JASON has placed considerable attention on events after the BW attack (after the "boom"). JASON repeatedly has made the argument that rapid detection and careful forensic analysis, even if it does not lead to attribution, shapes one's ability to gather and interpret intelligence information.

Over the years JASON has gone back and forth, considering both the pre-event intelligence and post-event forensics aspects of the timeline for a bioterrorism event. All the while the techniques for gathering intelligence and conducting forensic analysis have become more sophisticated, and global political and economic conditions have changed, as presumably have the underlying threats. The earliest JASON studies on bioterrorism predate 9/11 and focused mainly on what attacks might be possible given the principles of biomedicine and biotechnology and the potential capabilities of terrorist

groups (Koonin *et al.*, 2000). After 9/11 and the 2001 anthrax letters, the emphasis shifted to force protection and defense of the homeland in the face of an unspecified threat (Joyce *et al.*, 2001; Joyce *et al.*, 2002). It soon became apparent that it would not be possible to protect all critical resources from all possible bioterrorism threats. This conclusion was accompanied by the sober realization that the U.S. population itself constitutes the most effective sensor network for a bioterrorism attack – a distributed set of mobile, self-reporting individuals who themselves are the key assets to be protected. Subsequently the emphasis changed again, as JASON considered ways to compress the timeline for recognition of a bioterrorist attack and how to focus intelligence resources on the subset of potential threats that are the most plausible (Joyce *et al.*, 2005). Concern over the possibility of a rare but devastating bioterrorism event persisted, leading JASON to address potential doomsday threats such as synthetic smallpox and pandemic influenza (Stearns *et al.*, 2005; Block *et al.*, 2007). Most recently, JASON returned to the problem of forensics analysis for less exotic attacks, such as anthrax and tularemia, again with the goal of obtaining information to assist in intelligence gathering and threat assessment (Stearns *et al.*, 2008).

In light of the current study it is clear that what is needed is not only a link between forensics and intelligence, but also a bridge between common events of low consequence, occasional events of greater consequence, and the possibility of a rare and devastating attack. The question we come to is how best to characterize the base of the pyramid? Are there common events pertaining to bioterrorism, perhaps not involving bioterrorist attacks *per se*, that are part of the same probability distribution or can be related to the probability distribution that encompasses major bioterrorist attacks? For example, there is a growing "biohacker" community in the U.S. and abroad that

dabbles in recombinant DNA technology, genetic engineering, and synthetic biology in a manner akin to amateur pyrotechnics. Will this community become a source for, or an indicator of, groups having nefarious intent? Does activity in the biohacker community correlate with developments in the international terrorist community? Do small-scale events, whether successful or not, correlate in a statistical sense with a successful large-scale event? After all, the unsuccessful anthrax attack by Aum Shinriko was followed two years later by an attack with sarin gas in the Tokyo subway that killed 12 and injured thousands. It later was learned that Aum Shinriko attempted to obtain a virulent strain of Ebola virus, which might have been the basis for a large-scale event. Can one relate terrorist activities other than chemical and biological terrorism, such as planting IEDs or conducting suicide bombings, to the probability distribution that encompasses a major bioterrorism attack?

In order to develop a framework for relating actual low-level events to as-yet-unseen high-level events, it will be useful to characterize a middle ground along the same axis. For example, one could compile information pertaining to bioterrorist activity of any kind, including hoaxes, expressions of intent and the acquisition of restricted materials, and attempt to correlate these data with failed attempts and successful small-scale attacks. This information may bridge the gap to more infrequent events on a somewhat larger scale, which in turn may indicate the probability of a rare and more devastating event.

There often is a presumption that terrorist groups will make every attempt to conceal their activities with regard to bioterrorism. However, as was the case in Iraq under Saddam Hussein, there may be incentives for an organization to overstate it's capabilities for either internal or external pur-

poses. Furthermore, as discussed in Section 4.1.1, observation may trigger active deception, which could understate, overstate, or otherwise distort the true signal. Such distortions will degrade the predictive value of a statistical model, and may result in systematic errors that prevent meaningful extrapolation from common to rare events. This distortion can be mitigated by focusing on the middle ground, which links a broad range of observables to a modest number of *bona fide* events.

There is little doubt that a validated statistical model which links a rich database to a continuum of progressively rarer events would have utility in shaping one's posture with regard to the bioterrorism threat. Especially if that model were coupled to sensitivity analysis, it would focus intelligence gathering and forensic studies on those areas that would strengthen the model, and would guide the allocation of resources for asset protection and rapid response toward options that are expected to provide the greatest return on investment. It is likely, but needs to be demonstrated through the statistical model, that data gathering itself will have a high return on investment because it will enable the model to be refined, which will allow one to make ever smarter investments.

A APPENDIX: Black Swans

We frequently encountered references to "Black Swans" in our study. The Black Swan metaphor was popularized by a recent book *The Black Swan: The Impact of the Highly Improbable* by Nassim Taleb [39]. The metaphor has clearly had great impact on how people are thinking about rare events, so we considered Taleb's argument carefully.

Taleb's argument is that many high-magnitude rare events (Black Swans) are fundamentally unpredictable, and that efforts to predict them are futile, dangerous, and even intellectually fraudulent – particularly when using statistical models based on past observations, and especially if the statistical model assumes Gaussian variance around some mean event size. His term "Black Swan" is a reference to a (supposed) European belief that all swans were white until black swans were discovered in Western Australia in the 1700's. Black swans came to be used in philosophy as an example of a logical failure of inductive reasoning – that is, just because all previously observed swans are white does not necessarily mean that the next swan will be white.[7]

Taleb makes important points but carries his argument too far. It is unfortunately true that some mathematical models for risk forecasting assume Gaussian variance despite substantial countervailing real-world evidence. Taleb's criticism of economic risk models (such as the Black-Scholes-Merton options pricing model, which assumes that market prices make Gaussian-distributed moves) is particularly poignant given the world's current eco-

[7]This old philosophical argument is a little hard to digest from a probabilistic inference perspective. If we draw N white balls from an urn that might or might not contain black balls too, and if we assume we have no informative prior knowledge about the probability of white versus black balls might be, then the mean posterior estimate of the probability that the next ball drawn will be white is $\frac{N+1}{N+2}$, not one.

nomic situation. But though this is damning criticism of specific models' faulty assumptions, it is not a damning general criticism of the careful use of modeling to help predict risk.

Even in the original case of actual black swans, anyone sensible would have hesitated to extrapolate an observation of N white European swans to a prediction of zero black Western Australian swans. This obviously makes an assumption that Australian swans will be "drawn from the same distribution" as European swans, i.e., that Australian species are expected to be the same as European species. Thinking even for a second about the geographic distribution of species (especially strange Australian species), it doesn't take a wildlife biologist to doubt this assumption. In fact, black swans are a different species specific to Western Australia, *Cygnus atratus*. The black swan metaphor itself fails one of our key tests of predictive modeling – we would have worried that the observed (European swan) data are *not* likely to be adequately representative of the future (Australian swan) observations we want to predict.

Taleb includes the 9/11 attacks as an example of a "Black Swan" event. If we assumed that the magnitude of terrorist attacks were Gaussian-distributed, 9/11 would indeed appear to be an unforeseeably extreme outlier. In Section 3.3 we saw that the empirical data, though, says that the observed frequency/magnitude data may fit a power law distribution, and that 9/11 may not be a significant outlier in that distribution. Taleb is wrong that a terrorist event of 9/11's magnitude was fundamentally unforeseeable.

But Taleb is right that we should keep models' assumptions carefully in mind when we make any plans. In our power law model example, we already noted that we should hesitate before assuming that *all* future events will continue to be drawn from the same distribution. There is at least one class

of high-magnitude terrorist events that have not yet been observed but which clearly has a finite unknown probability: terrorist use of a nuclear weapon. Though we can draw a number of useful and relevant conclusions about the expected frequency and magnitude of future attacks using the same modalities as past attacks from the power law distribution, it would be dangerous to assume that the probability of extreme events is as low as the model predicts, because we (thankfully) have no data yet for the frequency/magnitude distribution of terrorist events using nuclear weapons. Taleb's recommendation of not overly relying on models based on past data, making subjective expert assessments of where additional unpredicted future risk might lurk, and making conservative investments to ameliorate the impact of unforeseen (even unforeseeable) negative "Black Swans" is fundamentally sound advice.

B APPENDIX: Rare Event Power Law Calculations

Section 3.3 presents some results based on the study in Clauset *e*t al˙ [31]. In this Appendix, we give some details behind those calculations. The reader is referred back to Section 3.3 for the discussion of appropriate use of such calculations.

B.1 Is 9/11 an Outlier?

We stated that the probability of an event of 9/11 magnitude or greater in the 1968-2006 period was about 23%, according to the fitted distribution in Clauset *e*t al. [31]. This may contradict one's intuition when reviewing the location of the 9/11 data point in Figure 5. However, it is important to remember this is not the tail of a Gaussian distribution, but of a heavy tailed distribution.

To calculate this probability, we need to consider the complementary cumulative distribution function for the power law, which is

$$P(X \geq x) = \frac{C}{(\alpha - 1)} x^{-(\alpha-1)},$$

where C is a normalization constant. Clauset *e*t al. [31] state $\alpha = 2.38$ but they don't give C. Any point from their fitted line suffices to estimate C. Using the x-intercept of the graph, which appears to be at about $x = 1100, P(X \geq 1100) = 10^{-4}$, we solve for $C = 2.17$. Given C and α, the probability $P(X \geq x)$ for any x can be obtained.

For the 9/11 point, $x = 2749$,

$$\begin{aligned} P(X \geq 2749) &= \frac{2.17}{(2.38-1)} 2749^{-(2.38-1)} \\ &= 0.0000282. \end{aligned}$$

There are N=9101 events in the dataset. Therefore, the expected number of events with $X \geq x$ is

$$\begin{aligned} E(X \geq x) &= NP(X \geq x) \\ &= 9101(0.0000282) \\ &= 0.257. \end{aligned}$$

The probability of one or more events greater than 2749 killed in the dataset of 9101 events is then given by a Poisson distribution,

$$\begin{aligned} 1 - e^{-NP(X \geq x)} &= 1 - e^{-0.257} \\ &= 0.226. \end{aligned}$$

This, finally, is the relevant probability. So, given the power law fit in [31], in the 9101 events over 38.5 years, the estimated probability of seeing one or more events of 9/11 New York magnitude is about 23%.

One way to double check this calculation is the following. A perfect fit would have $x \cong 1100$ as the maximal event. But the 9/11 New York point is at $x = 2749$, $\frac{2749}{1100} = 2.5$-fold higher than expected. In a power law complementary cumulative distribution function, for every multiplicative factor m that you increase x by, the probability $P(X \geq x)$ decreases by a factor of $m^{-(\alpha-1)}$. So, if $\alpha = 2.38$, a 2.5-fold increase in x reduces $P(X \geq x)$ by a factor of only about 0.28. This means that the point $x = 2749$ is only four-fold less likely than if it were perfectly on the power law line, not enough to consider it to be a significant outlier.

B.2 Odds of 9/11 Scale Event in Next Decade

To compute the change of a 9/11 scale event in the next decade, we again use the complementary cumulative distribution function $P(X \geq x)$ for the power law, with $\alpha = 2.38$ and $C = 2.1$. Assuming 2400 events per decade, we calculate the expected number of rare events of at least 9/11 magnitude occurring in the next ten years, as:

$$\begin{aligned} E(X \geq x) &= 2400 \times P(X \geq 2749) \\ &= 0.0676. \end{aligned}$$

The probability of one or more events greater than 2749 in the next decade is

$$1 - e^{-.0676} = .0654.$$

This suggests that another 9/11-scale event in the world is unlikely but not improbable in the next ten years, with a probability of about 7%.

C APPENDIX: Technical Note on Entropy

Imagine we have a discrete random variable with K possible values; as already mentioned one example might be the probability that a biological terror incident will originate from a group centered in country k, where k has a value between 1 and K. We have some observation of such incidents but not enough to trust that the simplest estimate $p_k = n_k/N$ (n_k = # of observed incidents from country k, N = total number of recorded incidents) is accurate. In particular, there may be many bins that have $n_k = 0$, but it might be foolish to conclude that all these probabilities are actually zero if the number of data points is small. The basic idea that will emerge from the mathematics is the importance of bins with $n_k > 1$; these can be called coincidences [1, 47]. Imagine there were no coincidences at all. Then, a reasonable assumption would be that the distribution is completely uniform and that some bins are at zero occurrences simply because $N < K$.

One way to characterize this distribution is by considering the entropy

$$S \equiv -\Sigma\, p_k \log_2 \rho_k.$$

For the uniform case, $p_k = \frac{1}{k}$, $S_{\max} = \log_2 K$. This, of course,is precisely the number of bits needed to completely specify which country was at fault for any particular instant, in the absence of any useful *a priori* information. Conversely, imagine that almost all observed events have occurred in a few bins. This gives much more confidence that the distribution really is peaked at a few hot spots and that the entropy is significantly less than $S_{\max}$. A natural definition of predictability [7] is just $S_{\max} - S$. Estimating the entropy from the data is of high priority.

Now for the mathematics. We will follow the discussion of Nemenman [1]

who developed these ideas for neural information processing, but which we feel have more general applicability. The staring point is use of the Bayesian method to estimate the

$$P(\vec{p};\vec{n}) = \frac{\prod_{i=1}^{K}\ p_i^{n_i}\ P_0(\vec{p})}{\int_0^1\ d^K p\ \prod_{i=1}^{K}\ p_i^{n_i}\ P_0\ (\vec{p})}.$$

The distributed $P_0(\vec{p})$ is the prior distribution placed on the probability vector $\vec{p}$. Here, $\vec{n}$ is the vector of observations which obeys $\sum_{i=1}^{k}\ n_i = N$. The left-hand side of the equation is our *a posteriori* estimate given the observations. A standard choice for the prior distribution is

$$P_\beta(\vec{p} = \frac{1}{Z\ (\beta)}\ \delta\left(1 - \sum_i^K p_i\right)\ \prod_{i=1}^{K})p_i^{\beta-1},$$

where the normalization factor $Z(\beta) = \Gamma^K(\beta)/\Gamma(K\beta)$. This has the virtue that the mean occupancy of each bin, as determined by the a posterior probability distribution, is just

$$< p_i >= \frac{n_i + \beta}{N + K\beta}.$$

This is similar to artificially adding observations of number β to each bin, This general notion subsumes specific choices in the literature; $\beta = 1$ (introduced originally by Laplace [48]), $\beta = 1/2$ (due to Jeffrey's [49]) and $\beta = 1/k$ (due to Schurmann and Grassberger [50]).

Nemenman's basic insight is that this usual procedure is not capable of giving a good estimate for the entropy. This is because the *a priori* distribution of the entropy for any fixed value of β is highly peaked in the large K limit. Thus, the prior distribution imposes a *pre-determined* entropy value on the final inferred result and does not allow for a fair estimation based on the data. In particular, the entropy is close to the maximal entropy for any finite value of β in the large K limit. To fix this, he generalizes this

prior distribution to a Dirchelet mixture of priors. He defines the weighted distribution

$$\tilde{P}_\beta = \frac{1}{Z}\,\delta(1-\sum\,p_i)\,\prod_{i=1}^{K}\,p_i^{\beta-1}\,\frac{d\xi}{d\beta},$$

where $\xi(\beta) = \psi_0(k\beta+1) - \psi_0\ (\beta+1)$ and where the function ψ_0 is the logarithmic derivative of the gamma function $\psi_0(x) \equiv \ \frac{d}{dx}\ \ln\Gamma(x)$. The instruction now is to define the overall probability distribution as an integral over β with this distribution. It can be shown that this choice leads to an almost uniform distribution of *a priori* entropy probability (between 0 and $S_{\max}$). Z is now an overall normalization factor.

The basic approach from here on in requires determining the value of β which dominates the *a posterioi* probability distribution. This leads [1] to the basic equation

$$\frac{1}{K}\sum_{i;n_i>0}\psi_0(n_i+\beta^*) - \frac{K_1}{K}\psi_0(\beta^*) - \psi_0(\beta^*K+N) = 0.$$

Here K_1 is the number of bins that have one or more counts. This can be solved to determine β^*, which then yields an estimate for the entropy $< S >= \xi(\beta^*)$. At the same time, we can get an estimate for the tail of the distribution (i.e., the probability of the rarest of the rare events)

$$P_i \sim \left(\frac{\beta\ B(\beta,\ k\beta-\beta)(k-i+1)}{k}\right)^{1/\beta}\ \ k-i << k,$$

where B is the incomplete B β function.

Laying aside all the details, the basic structure of the solution depends on the difference between K_1 and N. If $K_1 = N$, i.e., there are no coincidences at all, the solution predicts a finite β^* and $S \approx S_{\max}$. If $K_1 < N$, a non-trivial solution exists with the selected value of β decreasing as $N - K_1$ decreases, leading to a significant lower entropy estimate and a concomitantly higher predictability.

References

[1] Nemenman, I. (2002) Inference of Entropies of Discrete Random Variable with Unknown Cardinality. arXiv:physics/0207009v1 [physics.data-an].

[2] Todd, V., Cabayan, H., Ackerman, G., Asal, V., Brothers, A., Clark B., Connors, F., Diaz, J., Hunt, C., Kuznar, L., Nolte, P., Numerich, S., Pattipati, K., Popp, R., Rayner, M., Rethemeyer, K., Shellman, S., and Snead, E. (2008). Strategic Multi-layer Assessment of a Proposed Nationally Federated Capability to Perform Joint Intelligence Preparation of the Operational Environmnet Against the Potential Use of Weapons of Mass Destruction by Violent Non-State Actors. JIPOE SMA Interim Report, Alexander, VA.

[3] Cabayan, H. (2008) WMD-Terrorism (WMD-T) Joint Intelligence Preparation of the Environment (JIPOE) Concept Development Effort. Presentation presented to the Senior Review Team, Department of Defense: JS, OSD/DDRE, STRATCOM/GISC, Department of Homeland Security, Department of Energy, and ARGUS.

[4] Cabayan, H. (2009). Overview of Problem and What Led DoD to this Study. Presentation presented to JASON 2009 Summer Study Group.

[5] National Academy of Sciences (2003). *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences.* Committee on Responsibilities of Authorship in the Biological Sciences, National Academies Press, Washington D.C.

[6] Houghton, B. (2008) Terrorism Knowledge Base: A Eulogy (2004-2008). *Perspectives on Terrorism*, **11(7)**, 18-19.

[7] Bialek, W., Nemenman, I., and Tishby, N. (2001) Predictability, Complexity, and Learning. *Journal of Neural Computation*, **13**, 2409-2463.

[8] http://wits.nctc.gov/.

[9] http://www.rand.org/ise/projects/terrorismdatabase/.

[10] http://www.start.umd.edu/gtd/.

[11] http://dvn.iq.harvard.edu/dvn/dv/nds.

[12] http://www.nndb.com/.

[13] Kohut, A., Wike, R., Carriere-Kretschmer, E., and Holzwart, K. (2008) Unfavourable views of Jews and Muslims on the increase in Europe. Pew Global Attitudes Project Report, Washington, D.C..

[14] World Opinion http://www.worldpublicopinion.org/.

[15] http://www.gallup.com/consulting/worldpoll/.

[16] http://pewglobal.org/.

[17] http://www.norc.org/GSS/GSS+Resources.htm.

[18] Newman, M. (2006) Power Laws, Pareto Distributions and Zipf's Law. arXiv:cond-mat/0412004v3 [cond-mat.stat-mech].

[19] SMA (2008) *Anticipating Rare Events: Can Acts of Terror, Use of Weapons of Mass Destruction or Other Hight Profiel Acts be Anticipated? A Scientific Perspective on Problems, Pitfalls and Prospective Solutions* Edited by Nancy Chesser, Washington D.C.

[20] SMA (2009a) *Collaboration in the National Security Arena: Myths and Reality - What Science and Experience Can Contribute to its Success.* Edited by Jennifer O'Connor, Elisa Bienenstock, Robert Briggs, Carl Dodd, Carl Hunt, Kathleen Kiernan, Joan McIntyre, Randy Pherson, and Tom Rieger, Washington D.C...

[21] SMA (2009b) *From the Mind to the Feet: Assessing the Perception-to-Intent-to-Action Dynamic.* Edited by Larry Kuznar and Allison Astorino-Courtois, Washington D.C.

[22] Kiernan, K., and Mabrey, D. (2009) From Shoe Leather to Satellites. In *From the Mind to the Feet: Assessing the Perception-to-Intent-to-Action Dynamic.* Edited by Larry Kuznar and Allison Astorino-Courtois, Washington D.C., 15-18.

[23] Schaub, G. (2009) Gauging the Intent of Nation-States and Non-State Actors: An Operators Perspective. *From the Mind to the Feet: Assessing the Perception-to-Intent-to-Action Dynamic.,* Edited by Larry Kuznar and Allison Astorino-Courtois, Washington D.C., 26-37.

[24] Yarhi-Milo, K. (2009) Intent form an International Politics Perspective: Decision-Makers, Intelligence Communities and Assessment of Adversary's Intentions. *From the Mind to the Feet: Assessing the Perception-to-Intent-to-Action Dynamic.* Edited by Larry Kuznar and Allison Astorino-Courtois, Washington D.C., 59-67.

[25] Rieger, T. (2009) Survey Science: Desperate Measures, Different Types of Violence, Motivations and Impact on Stability. *From the Mind to the Feet: Assessing the Perception-to-Intent-to-Action Dynamic.* Edited by Larry Kuznar and Allison Astorino-Courtois, Washington D.C., 84-90.

[26] RAND National Defense Research Institute (2009). *Social Science for Counterterrorism*. Edited by Paul Davis and Kim Cragin. RAND Corporation, Santa Monica, CA.

[27] http://earthquake.usgs.gov/research/parkfield/scholz.html.

[28] Saegusa, A. (1999) China Clamps Down on Inaccurate Warnings. *Nature*, **397**, 284.

[29] http://www.usgs.gov/science/science.php?term=302,
http://earthquake.usgs.gov/research/hazmaps/design/

[30] Christensen, K., Danon, L., Scanlon, T., and Bak, P. (2002) Unified Scaling Law for Earthquakes. *Proceedings of the National Academy of Sciences*, **99**, 2509-2513.

[31] Clauset, A., Young, M., and Gleditsch, K. (2008) On the Frequency of Severe Terrorist Events. *Journal of Conflict Resolution*, **51**, 58-87.

[32] Heinrich, H. (1950) *Industrial Accident Prevention*. McGraw Hill, New York, NY.

[33] http://secondlife.com/.

[34] http://lindenlab.com/.

[35] Curis, K. (2001) Multinational Information Sharing and Collaborative Planning Limited Objective Experiments. MITRE Corporation report, McLean, VA.

[36] http://grants.nih.gov/grants/guide/pa-files/PAR-07-344.html.

[37] http://www.militaryspot.com/news/item/joint_forces_command_interagency_experiment_prepares_for_crisis/.

[38] http://www.nowpublic.com/terrorism_is_temporary_in_second_life.

[39] Taleb,. N. (2007) *The Black Swan: The Impact of the Highly Improbable*. Random House, New York, NY.

[40] http://www.start.umd.edu/start/research/projects/project.asp?id=58.

[41] http://www.haaretz.com/hasen/spages/1087126.html.

[42] http://www.jihadwatch.org/archives/024803.php.

[43] http://www.haaretz.com/hasen/spages/1087126.html.

[44] Kull, S. (2007) Muslim Public Opinion on U.S. Policy, Attacks on Civilians and Al Qaeda. World Public Opinion.org project report, University of Maryland, VA.

[45] Esposito, J., and Mogahed, D. (2008) *Who Speaks For Islam? What a Billion Muslims Really Think*. Gallup Press, New York, NY.

[46] http://www.start.umd.edu/start/publications/research_briefs/20061120_pipa.pdf.

[47] Ma S. (1981) Estimating Entropy Rates with Bayesian Confidence Intervals. *Journal of Statistical Physics*, **26**, 221-240.

[48] Wolpert, D., and Wolf D. (1995) Estimating Functions of Probability Distributions from a Finite Set of Samples. *Physical Reviews E*, **52**, 6841-6854.

[49] Jeffreys, H. (1946) An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society (London) A*,**186**, 453-461.

[50] Schurmann T. and Grassberger, P. (1996) Entropy estimation of symbol sequences. *Chaos*, **6**, 414-427.

[51] Myerson, R. (1997), *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

[52] Block, S., Joyce. G., Dyson, F., Haussler, D., Levine, H., Nelson, D., Press, W., Schwitters, R., Stearns, T., and Weinberger, P. (2007) Synthetic viruses. JASON Report JSR-07-508, The MITRE Corporation, McLean, VA.

[53] Joyce, G., Schwitters, R., Gifford, D., Happer, W., Henderson, R., Koonin, S., Levine, H., Lewis, N., Weinberger, P., and Williams, E. (2001) Biosensing. JASON Report JSR-01-100, The MITRE Corporation, McLean, VA.

[54] Joyce, G., Abarbanel, H., Block, S., Drell, S., Dyson, F., Henderson, H., Koonin, S., Lewis, N., Schwitters, R., Weinberger, P., and Williams, E. (2002) Biodetection architectures. JASON Report JSR-02-330, The MITRE Corporation, McLean, VA.

[55] Joyce, G., Block, S., Brenner, M., Dyson, F., Grober, R., Happer, W., Haussler, D., Hemley, R., Hwa, T., Koonin, S., Levine, H., Muller, R., Nelson, D., Prentiss, M., Stearns, T., Weinberger, P., and Woodin, H. (2005) Emerging viruses. JASON Report JSR-05-502, The MITRE Corporation, McLean, VA.

[56] Keim, P., L. Smith, K.L., Keys, C. Takahashi, H., Kurata, T., and Kaufmann, A. (2001) Molecular investigation of the Aum Shinrikyo anthrax release in Kameido, Japan. J. Clin. Microbiol. 39, 45664567.

[57] Koonin, S., Abarbanel, H., Cornwall, J., Dimotakis, P., Drell, S., Dyson, F., Fortson, N., Garwin, R., Henderson, R., Jeanloz, R., Joyce, G., Katz, J., Lewis, N., Schwitters, R., Stubbs, C., Sullivan, J., Weinberger, P.,

and Young, C. (2000) Civilian biodefense. JASON Report JSR-99-105, The MITRE Corporation, McLean, VA.

[58] Stearns, T., Block, S., Dyson, F., Haussler, D., Hemley, R., Joyce, G., Levine, H., Nelson, D., Press, W., and Woodin, H. (2005) BioEngineering. JASON Report JSR-05-130, The MITRE Corporation, McLean, VA.

[59] Stearns, T., Block, S., Breaker, R., Brenner, M., Dyson, F., Joyce, G., Levine, H., Nelson, D., Weinberger, P., and Westervelt, R. (2008) Microbial forensics. JASON Report JSR-08-512, The MITRE Corporation, McLean, VA.