This report is part of the RAND Corporation research report series. RAND reports present research findings and objective analysis that address the challenges facing the public and private sectors. All RAND reports undergo rigorous peer review to ensure high standards for research quality and objectivity.

NATIONAL DEFENSE RESEARCH INSTITUTE

Using Behavioral Indicators to Help Detect Potential Violent Acts

A Review of the Science Base

Paul K. Davis, Walter L. Perry, Ryan Andrew Brown, Douglas Yeung, Parisa Roshan, Phoenix Voorhies

Prepared for the United States Navy
Approved for public release; distribution unlimited

The research described in this report was prepared for the United States Navy. The research was conducted within the RAND National Defense Research Institute, a federally funded research and development center sponsored by the Office of the Secretary of Defense, the Joint Staff, the Unified Combatant Commands, the Navy, the Marine Corps, the defense agencies, and the defense Intelligence Community under Contract W74V8H-06-0002.

Library of Congress Cataloging-in-Publication Data

Davis, Paul K., 1952-
Using behavioral indicators to help detect potential violent acts : a review of the science base / Paul K. Davis, Walter L. Perry, Ryan Andrew Brown, Douglas Yeung, Parisa Roshan, Phoenix Voorhies.
pages cm
Includes bibliographical references.
ISBN 978-0-8330-8092-9 (pbk. : alk. paper)
Terrorism—Prevention. 2. Behavioral assessment. 3. Psychology—Methodology. I. Title.

HV6431.D3268 2013
363.325'12--dc23

2013024014

The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Support RAND—make a tax-deductible charitable contribution at www.rand.org/giving/contribute.html

RAND® is a registered trademark

Cover photo by Karl Baron via flickr

RAND OFFICES
SANTA MONICA, CA • WASHINGTON, DC
PITTSBURGH, PA • NEW ORLEANS, LA • JACKSON, MS • BOSTON, MA
DOHA, QA • CAMBRIDGE, UK • BRUSSELS, BE
www.rand.org

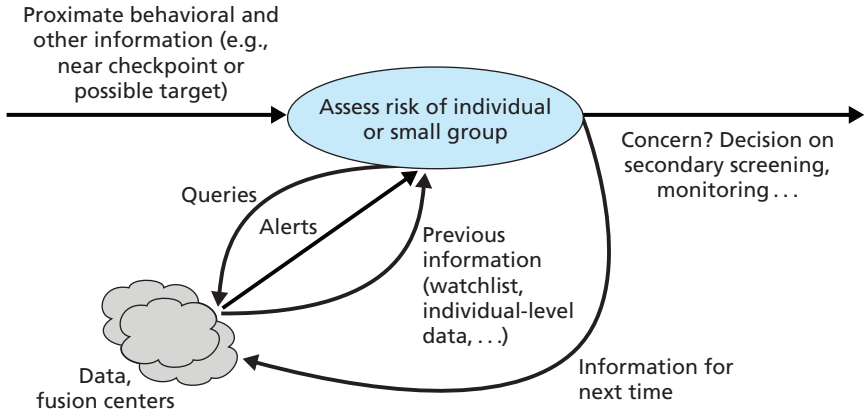
Summary

Government organizations have put substantial effort into detecting and thwarting terrorist and insurgent attacks by observing suspicious behaviors of individuals at transportation checkpoints and elsewhere. Related technologies and methodologies abound, but their volume and diversity has sometimes been overwhelming. Also, effectiveness claims sometimes lack a clear basis in science and technology. The RAND Corporation was asked to review the literature to characterize the base in behavioral sciences relevant to threat detection, in part to help set priorities for special attention and investment.

Purpose and Approach

Our study focused on the science base for using *new or nontraditional* technologies and methods to observe behaviors and how the data gathered from doing so might—especially when used with other information—help detect potential violent attacks, such as by suicide bombers or, as a very different example, insurgents laying improvised explosive devices (IEDs). Behavioral indicators may help identify individuals meriting additional observation in an operational context such as depicted in Figure S.1. For that context, security personnel at a checkpoint are assessing (blue oval) whether an individual poses some risk in the limited sense of meriting more extensive and perhaps aggressive screening, follow-up monitoring, or intercept. They obtain information directly, query databases and future versions of information-fusion centers ("pull"), and are automatically provided alerts and other data

Figure S.1
A Contextual View of the Detection Effort



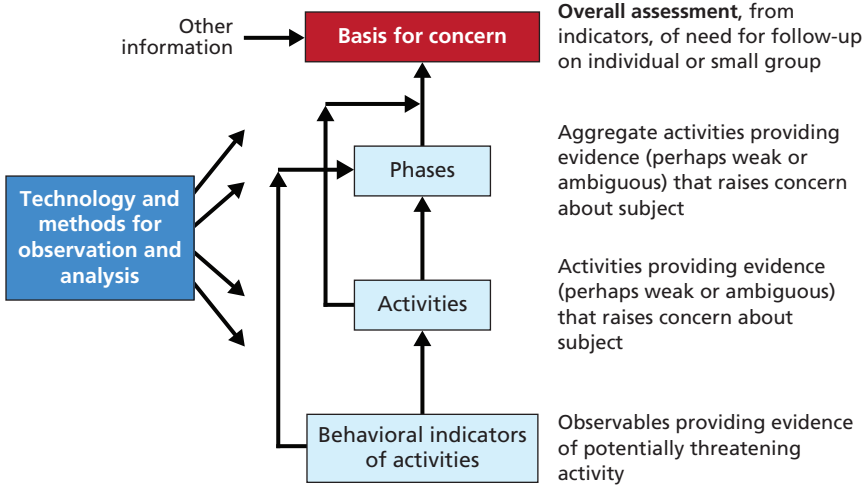
RAND *RR215-S.1*

("push"). They report information that can be used subsequently. In some cases, behaviors of a number of individuals over time might suggest a potential ongoing attack, even if the individuals are only pawns performing such narrow tasks as obtaining information.

Figure S.1 refers to "other information" (top left). Although our study is concerned with detecting imminent threats rather than gathering broad information for internal security or intelligence, such information—perhaps accumulated over years—can play an important role. Where might that information be found, how might it be structured, and what indicators might be involved? We focus on what may be possible technically, without analyzing tradeoffs with privacy and civil liberties. We do, however, note troublesome issues raised by the technology and methods, point readers to an in-depth study of such matters by the National Academy of Sciences, and suggest research on ways to mitigate the problems.

Figure S.2 shows relationships among our key constructs. A base of technology and methods (left) allows detecting behavioral indicators (bottom right). Moving upward, these give signals about activities, which are grouped into activity classes called phases. Analysis can then assess whether the totality of information (including nonbehavioral

Figure S.2
Relationships Among Constructs



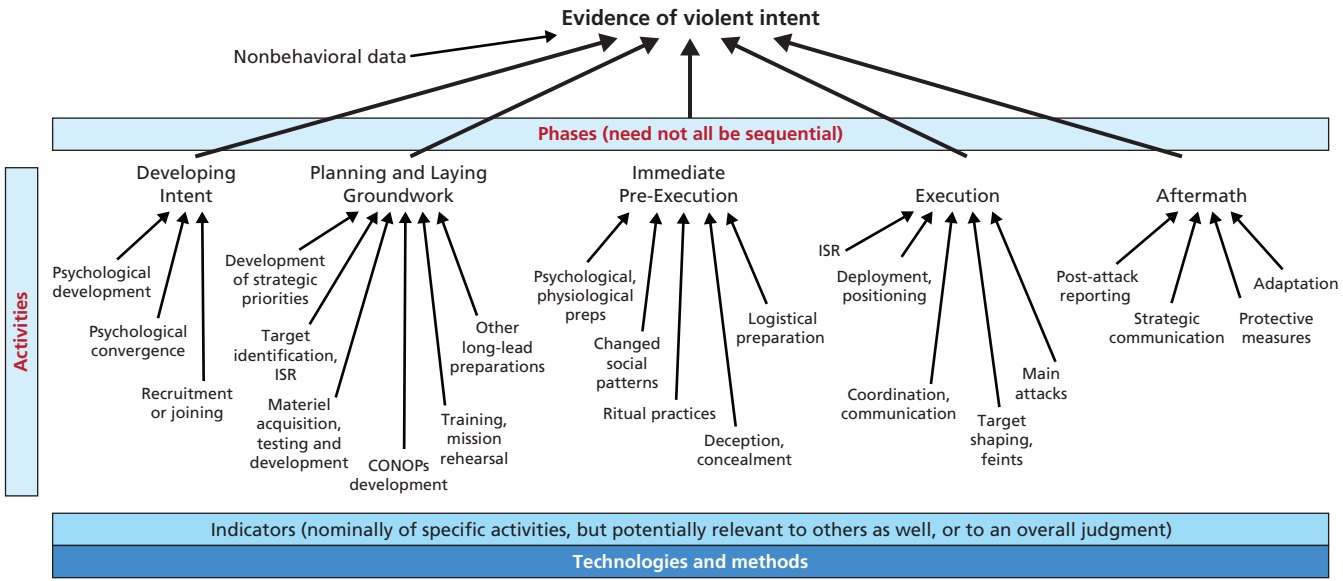
RAND *RR215-S.2*

information) adds up to a basis for concern justifying more screening, monitoring, precautionary defensive measures, or preemptive action. Since detecting a "basis for concern" (i.e., need for further checking) will probably have a high false alarm rate, a system using this approach must be efficient and reasonable if it is to be acceptable.

Figure S.3 is a conceptual model showing phases within which lower-level activities occur. The model merely identifies where to look for information. As indicated at the bottom of the figure, there are many possible indicators of the activities and a number of technologies and methods to use in observing them. The model is merely heuristic rather than a rigorous decomposition or timeline. Activities may be performed by multiple individuals, not occur, or occur in different order. Some activities could occur in more than one phase.

Figure S.4 uses the "Developing Intent" phase of Figure S.3 to illustrate how phases, activities, and indicators relate to technologies and methods. For each of three activities, Figure S.4 shows potential indicators. The lowest box shows some relevant technologies and methods. The Developing Intent phase is unusual in that it includes

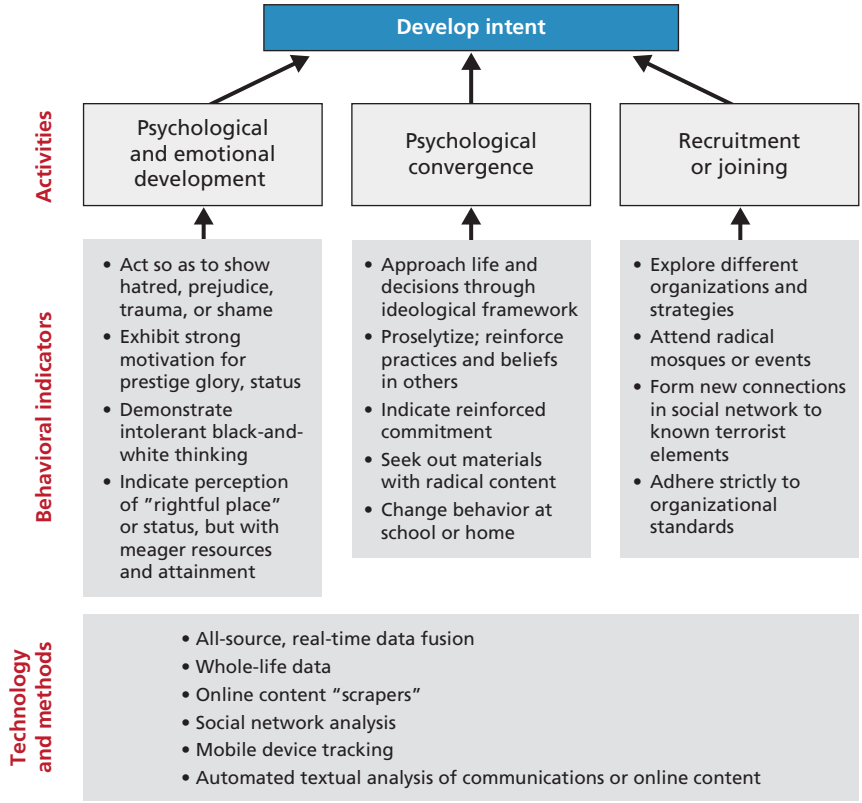
Figure S.3
Conceptual Model of Opportunities for Observation



NOTES: ISR = intelligence, surveillance, and reconnaissance. CONOPs = concept of operations. Indicator-phase connections are nominal. Some activities can occur in more than the phase indicated. Associations of activities with phases are nominal. Some activities can occur in additional phases.

RAND *RR215-S.3*

Figure S.4
Illustration of Methodology



RAND *RR215-S.4*

early-in-life activities, such as might be observed by parents, neighbors, teachers, physicians, local law enforcement, and others long before an individual becomes involved in anything violent. In the main text, we discuss the phases separately, identifying generic activities and potential indicators.

Technology and Methods

We found it useful to highlight technologies and methods in three cross-cutting classes of data: (1) communication patterns, (2) "pattern-of-life" data, and (3) data relating to body movement and physiological state. Most of the methods suffer from signal-to-noise problems and false alarms; some are vulnerable to countermeasures.

Communication Patterns

Communications occurs in, e.g., face-to-face meetings, Internet chat rooms, and cell phones. Large commercial and intelligence-sector investments have yielded techniques to monitor and analyze these communications, which we also treat in three groups: online communications and analysis, text analysis and natural-language processing, and speech analysis.

Online Communications. Using data collected by online-content "scrapers" and subsequent human efforts, it is sometimes possible to infer motivations, emotions, and ideologies from online statements and actions (e.g., as with the New York Police Department following Twitter posts such as "people are gonna die like Aurora" (referring to the July 20, 2012, movie-theater shooting in Aurora, Colorado). Real-time social-media search tools can help monitor and track discussions of potential targets. Keystroke loggers and other programs can reveal past and current searches for materiel; registrations or payments to training programs; location information; and information on past searches. Changes of activity may occur when terrorists "go dark" before an attack, when logistical preparations for an attack are intense, or when calls for vengeance arise after an event such as Osama bin Laden's killing.

This and all the methods include false alarms (e.g., users may have benign reasons for their purchases or mere curiosity as they investigate troublesome websites), low signal-to-noise ratio, and vulnerability to such counters as burying information amidst innocuous communication or using anonymous or false accounts.

Text Analysis and Natural-Language Processing. Techniques to classify texts and analyze content are fairly well developed, although less so

with respect to emotion and intent. Explicit content may include bragging, ideological statements, or admiration for terrorist leaders. Textual analysis of *style* can detect word-usage patterns associated with typical attacker motivations and such emotions as anger, humiliation, and shame. Style of communication does not depend on content, i.e., on specific topic, and can suggest relationships and status within a social network. Textual analysis of style has been used for, e.g., detecting corporate fraud, terrorist interrogations, and criminal testimony. Natural-language processing can analyze massive amounts of text about which little is known in advance, classifying documents to be analyzed further by subject-matter experts. Clustering methods can identify concepts and such topics as weapons, tactics, or targets. Such mathematical techniques as latent semantic indexing can help understand concepts and have the advantage of being language-independent. Machine translation can often turn foreign-language texts into something analyzable without foreign-language expertise or language-specialized software. Speech-recognition technology can greatly increase the amount of text available for text analysis. It can also help identify individuals.

A primary shortcoming is nonspecificity; that is, detected patterns (even if apparently threatening) may be unrelated to any imminent threat, and their interpretation often depends on cultural and individual idiosyncrasies. A shortcoming of the research base itself is that much linguistic-style analysis has been done only on archival data; more testing and validation is needed with "real-life" data sets. Top researchers caution against expecting highly reliable detections or interpretations and suggest the need for very large data sets that reveal many cultural and individual differences.

Speech Analysis of Content. Several robust indicators exist for connecting vocal content and narratives with lying and deception. These include the subject (1) distancing himself from untruthful statements by, e.g., using the third person or otherwise seeming less verbally involved; (2) issuing discrepant statements; (3) providing less detail; (4) exhibiting less logical structure and less subjectively plausible stories; (5) providing less context; and (6) making less spontaneous corrections or claiming lapses of memory.

This approach's primary shortcoming in assessing deceptive or hostile intent is that interpreting lexical and vocal indicators of lying and deception depends on context, individual variability, and appreciation of nonthreatening explanations. Optimally, analysis has data on the individual's speech in a normal nondeceptive/nonhostile state. Where this is infeasible, the potential increases markedly for failed detections and intolerably many false alarms. Table S.1 summarizes

Table S.1
Considerations and Caveats: Detection and Analysis of Communication Patterns

Domain	Status	Upside Potential	Measurement Requirements	Shortcomings and Vulnerabilities
Online communication and activities	Extensive collection and analysis occurs today for commercial and intelligence reasons. Technologies and methods for analyzing such online activities are still relatively unproven in either academic or operational settings.	Given trends, even more and varied interactions will be available for collection.	Tools already exist. However, challenges for dealing with massive volumes of noisy data are formidable.	Methods have not been well validated in academic or operational settings. Low signal-to-noise ratio. Effects of encryption, using "code," using anonymizers, or moving offline.
Text analysis and natural-language processing	A considerable research base exists with numerous past applications. Even natural-language processing can be highly accurate in specific experimental settings.	Using operational data to train and to create baselines could improve detection of deception, hostility, or extremist patterns. Natural-language techniques, given training sets, could quickly analyze large amounts of data.	Online text is naturally occurring and publicly accessible, requiring only passive collection. Active elicitation of text or oral statements is possible in some security contexts, such as checkpoints or interrogations.	Context and cultural dependence. Inadequate testing in operational settings. Need for substantial data.

Table S.1—Continued

Domain	Status	Upside Potential	Measurement Requirements	Shortcomings and Vulnerabilities
Speech analysis: lexical and vocal cues	This has been validated in laboratory settings, including those specific to counterterrorism.	Advances in protocols for rapid assessment of speech patterns and content would have wide applicability for screening, checkpoint, or other situations involving conversations with security personnel.	Such analysis currently requires skilled security personnel asking questions and making judgments.	Physiological drivers, such as anxiety and changes in vocal tone, are individual-dependent. May be subject to counters, especially if criteria for judging are known.

results of our review for the assessment of communication patterns and content.

Pattern-of-Life Data

It is possible to analyze patterns of communication, travel, purchasing, and other matters using existing records and databases (many held by private industry). We discuss mobile-device tracking, using existing records, and machine learning for pattern detection. These raise profound social questions about what kind of data can and should be collected and analyzed.

Mobile-Device Tracking. Ubiquitous mobile devices provide a wealth of data on personal information, social relationships and networks, locations, patterns of movement and interactions, preferences, political opinions, the spread of information, and patterns of how opinions and preferences change. Also, mobile-device usage is related to the "Big Five" personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism).

One shortcoming of such data is that much social networking through mobile devices is increasingly "muddy" and in many cases divorced from both intent to meet and intent to act in the offline "real

world." This complicates inference-making about patterns of communication and their connection to actual threat. And, of course, people can go offline.

Existing Records. It is sometimes possible to develop individual profiles from information about, e.g., experiences, behaviors, and relationships over time, and to provide context for assessing other incoming data. The data could come from school records, criminal records, interrogation reports, and so forth. Additionally, surveillance cameras are now common in public and business settings, allowing for the possibility of tracking individual patterns-of-life. Integrating such data requires analytic techniques, including those for all-source, real-time big-data fusion. Related analytic tools are increasingly available from such providers of cloud computing as Google and Amazon and social-media companies.

The shortcomings include, of course, the administrative, jurisdictional, legal, and database challenges of extracting and combining data across multiple sources and owners within and outside the United States. Accuracy matters for this type of analysis, whereas commercial applications often do not require high accuracy to improve the targeting of marketing efforts.

Machine Learning and Big-Data Analysis. Given the sheer magnitude of data, it is increasingly important to analyze information without the benefit of prior hypotheses or known points of comparison. "Supervised" machine-learning techniques use known data sets to train the algorithms, which can then classify data, identify relationships, and discover concepts. "Unsupervised" learning proceeds without the aid of such known prior knowledge. It seeks to find structure in the unlabeled data set. For example, researchers have used thousands of YouTube images for unsupervised detection of high-level features such as faces. Potentially, such techniques could recognize images suggesting imminent threat. Such machine-learning techniques have been applied to uncover fraud, to recognize deception in computer-mediated communication, and for predictive policing. Artificial neural-network models are promising and can be applied in real time. Video or image analysis and machine-learning techniques could be employed to find,

for example, such activities as shaving heads and prayer activities in martyrdom videos.

One shortcoming is that machine-learning techniques often require a large amount of data. At least in the public domain, sufficiently large databases of violent attacks and other events do not exist for topics such as terrorism. One innovative method for obtaining large, labeled data sets is to "crowd-source" collecting and labeling individual pieces of information.

Table S.2, which is analogous with Table S.1, is our assessment of the various approaches focused on records-based whole-life information.

Indicators from Physical Movement and Physiology

Behavioral science has identified many nonverbal behaviors that are statistically associated with emotional and psychological state and with deception or violent intent. These can be roughly categorized into (1) kinetics (including gross motor movements) and (2) observation of physiological state. As discussed in the main text, *many* findings are controversial among scientists, and between scientists and operators, but our summary assessment follows.

Kinetics and Gross Movement. Existing technology can collect data on kinematic patterns (movement). Surveillance and reconnaissance platforms (e.g., tower cameras or drone systems) can monitor individuals as they maneuver before an attack. Video systems can view individuals before attacks and collect information on individuals who frequent potential attack sites, providing a baseline for identifying individuals engaged in pre-execution activities. For example, the gait of people who may be carrying weighted objects, such as IEDs, may be compared against baseline "gait signatures." Existing recordings of terrorism incidents can provide data for setting parameters on new analysis tools. The Defense Advanced Research Projects Agency (DARPA) has funded biometric technologies for identification at a distance and for early warning. Another approach seeks to automate recognition of potentially threatening body postures or poses.

Incorporating emotion into machine-learning methods may increase their future utility. To do so, "affective computing" may need

Table S.2
Considerations and Caveats for Pattern-of-Life Data

Domain	Status	Upside Potential	Measurement Requirements	Shortcomings and Vulnerabilities
Mobile-device tracking	Algorithms to predict individual movement patterns, preferences, etc., have been developed and validated in laboratory and experimental settings, but can benefit from more naturalistic validation.	Mobile devices will continue to add connectivity features that enable tracking (e.g., location and motion sensors, Near Field Communication chips).	Mobile device tracking may require device-owner permissions or cooperation of communications network providers.	Not traveling with or turning off device will defeat methods based on mobile-device whereabouts. Mobile-to-mobile communication is often divorced from "real-life" behaviors and intent.
Pattern-of-life data	Validating techniques to analyze large amounts of pattern-of-life data may be difficult in academic settings. Commercial data sets and analytic tools are increasingly available.	Pattern-of-life data may allow integrating disparate data types to build fuller behavioral profiles on individuals of interest. Accessing and integrating data is an issue.	Measurement does not require active or voluntary consent. However, access to various databases held by commercial or private sources may be necessary.	Pattern-of-life data may be vulnerable to "cover" activities and behaviors. Databases and algorithms for detecting threatening patterns are in early development.
Machine learning and big-data analysis	Machine-learning techniques have been extensively used and validated in experimental and some applied settings. Such techniques have been used in national security and law enforcement.	Machine-learning and big-data analysis may "discover" unknown patterns or activities hidden in large amounts of data, but massive amounts of data are needed for training.	Measurement does not require active or voluntary consent. A large amount of data or a strong hypothesis regarding relevant activity is required.	Learning techniques are probabilistic and vulnerable to noisy data. Current systems do not understand how to associate behaviors of multiple threatening individuals.

to select from various psychology and neuroscience findings and theories of emotion (e.g., "appraisal models"). Often, subsystems of monitoring and interpretation of stimuli can be computationally modeled. Improvements are possible when distinguishing between emotional states that differ in arousal, such as anger and sadness. Methods being developed to analyze gait signatures could be applied to such existing commercial technologies as cameras used for the Microsoft Kinect and Nintendo Wii game systems' motion-capture capability.

Human observers and analysts may also be employed, but results depend on such factors as training, individual talent, and observer bias. Detecting deceptive movements is easier for people experienced in employing the same deceptive movement patterns. People are best able to detect emotions associated with gait when the human walkers are expressing anger. Inference from merely a single stride can be highly accurate, suggesting that gait can be used to recognize affect. Performance varies by individual, and women may be better than men at determining actions and recognizing emotions from movements such as walking.

Analysis of kinetics and gross motor movements should apply to a wide variety of security contexts, although validation in naturalistic settings is needed and, as often occurs in looking for behavioral indicators, the indicators may arise for benign reasons, such as people being anxious at security screenings or checkpoints.

One challenge for gait analysis is that current detection systems and protocols are often built using simulated behaviors (e.g., with actors). More naturalistic (real-world) observations are needed.

Physiological State and Reactions. Observing physiological state holds promise for detecting deception and other behaviors. We touch upon polygraph testing, other measures of peripheral nervous system response, electroencephalograms (EEGs), vocal stress, and facial-expression analysis.

Polygraph testing has long been employed and found useful as *part* of an investigatory process (particularly because people often "open up" in the process), but is not by itself reliable. A great divide exists between the bulk of the academic community, who remain quite skeptical, and the community of "operators," who insist on its useful-

ness as one tool in a process. Newer approaches using some of the same physiological signals as in polygraphs (heart rate, blood pressure, etc.) are in development with respect to detection of potential deception or hostile intent.

New technologies using electroencephalograms (EEGs) allow some physiological features to be observed without "wiring up" individuals, sometimes at a distance, and sometimes covertly or surreptitiously, as with using heat-sensitive cameras to detect capillary dilation and blood flow to the face and head. There is some evidence of unique value in indicating deception or imminent action by an individual if baseline information is available for that specific individual ahead of time or if credible intelligence about a possible attack is available. Most of the technologies are in a relatively early stage of development, but there does seem to be potential. Measurement of physiological signals closer to the central nervous system (i.e., the brain) holds the most promise for detecting guilt and behavioral intent.

Evidence of vocal tension and higher vocal frequency may also be predictors of stress and deception, and a few observable aspects of speech are much more difficult for an individual to *control* than other indicators of deception, but countermeasures that obscure differences from the baseline of normalcy are definitely feasible.

Humans appear to share universal facial expressions indicative of underlying emotional and motivational states. Cultural differences seem to affect only secondary aspects of facial expressions. The seven fundamental emotions—anger, disgust, fear, happiness, sadness, surprise, and contempt—are displayed on the face with some fundamental features that are generally recognizable on all humans (barring neurological impairment). For the purposes of detecting pre-incident indicators, the most promising domain of facial expression analysis involves facial micro-expressions—involuntary expressions of emotion appearing for milliseconds despite best efforts to dampen or hide them. Whether the relevant behavior is smuggling weapons, traveling on forged documents, or hiding anger or anxiety near security officials, facial micro-expressions can be important indicators of deception or some kind of mal-intent.

At least currently, the two primary problems with using physiological indicators are (1) nonspecificity (the indicators may stem from many causes, most of them benign) and (2) individual differences (the observables that indicate attack or deception differ markedly across individuals, which requires establishing sound individual-centered baselines). Countermeasures are a problem with polygraphs, but perhaps less so with EEG methods. Even with polygraphs, empirical results have varied. Some drugs, for example, have not reduced detection rates as expected, but physical training can be effective as a countermeasure. Controlling vocal stress indicators is difficult, but countermeasures can obscure distinctions between baseline and stressed behavior. Facial expressions suffer from the same problems of nonspecificity, but they have the advantage of being more closely linked to motivational state and intent than are other physiological signals. Individual differences are also important: A psychopathic attacker, for example, might be more likely to show micro-expressions of "duper's delight" while passing through a checkpoint undetected, while a nonpsychopathic attacker might instead show micro-expressions of fear (as would a perfectly harmless nervous traveler).

While the link between micro-expressions and deception is well evidenced, utility in security-related settings is another matter. Coding emotional expressions currently involves hours of labor to analyze seconds of data, making this technique unsuitable for use in real time at checkpoints or other screening areas. However, a training system appears to increase the capacity of individuals to detect facial expressions and micro-expressions with demonstrated evidence of effectiveness in clinical populations.

Recognition of emotional expressions based on automated algorithms and computation is still in its infancy, but this is an active field of development, and improved algorithms are likely to yield greater accuracy and robustness. Furthermore, as with many pre-incident indicators of attack, emotion-recognition algorithms that fuse multiple parameters seem to perform much better than inferring emotional state simply from facial expressions alone.

Of course, checkpoints or other security environments are dynamic locations where it is difficult to capture high-resolution video

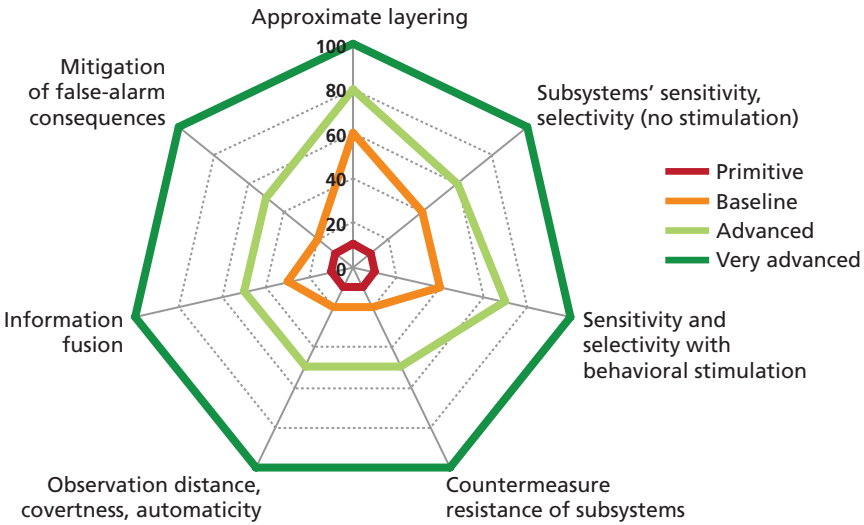
(or audio or physiological data) for individuals, but such detailed information is often necessary for effectiveness.

Table S.3 is our assessment of how the approaches based on detecting intent from physiological indicators stand in terms of maturity, potential, measurability, and vulnerability to countermeasures.

Cross-Cutting Themes

A number of cross-cutting issues arose in our review. These suggest a notional framework for thinking about detection systems. Although the relevant metrics have by no means been defined as yet, much less metrics that take into account cost-effectiveness, a goal for future analysis might be to place something like the framework shown in Figure S.5 on a solid scientific and analytic basis. Although it is surely not yet "right" or well defined, this framework conveys a sense of what is needed for sounder discussion. Further, despite its shortcomings, we

Figure S.5
A Notional Framework for Characterizing an Overall System



RAND *RR215-S.5*

Table S.3
Detecting Hostility or Deception from Movement Physiology and Movement

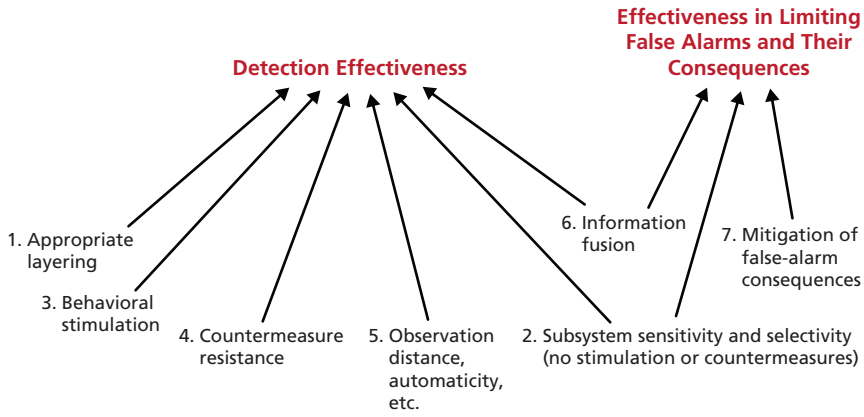
Domain	Status	Upside Potential	Measurement Requirements	Shortcomings and Vulnerabilities
Kinetics and gross motor movement	Indicators have been validated for human observation and automated analysis in laboratory and experimental settings, including some operational settings (e.g., for gait of individuals carrying weighted objects).	Gross motor movements may reveal action, intent, or deception. On-foot motions may be unavoidable in such proximal security settings as checkpoints. Gross motor movement may be passively observed, but also actively elicited.	Some security contexts may not allow for sufficient physical movement to be interpretable (e.g., interrogation).	Masking with deceptive movements. Sensitivity to context and individual differences. Nonspecificity: triggering by diverse emotions and motivations.
Physiological state and reactions	Indicators have been validated in laboratory and experimental settings, with some experimental paradigms simulating elements of counterterrorism and some (facial) cutting across culture. In some cases (e.g., voice stress and facial indicators), automated recognition shows potential but currently has high error rates.	Internal physiological reactions are relatively automatic and difficult to control (e.g., micro tremors in speech or micro facial expressions). Probing of various sorts (even seemingly random conversations) can trigger reactions. Certain elements of facial expression are very difficult to alter voluntarily, including micro-expressions.	Currently, measurement requires direct application of sensors or the physical observation of, e.g., facial flushing, sweating. Some (e.g., facial) require lighting and proximity with currently painstaking coding feasible only for high-value interrogations. Success requires exceptional "natural" talent or training, but limited available data suggests training is effective. Measurements are most valuable when comparing against an individual's baseline, which is only feasible in voluntary monitoring or interrogation context.	Differences across contexts and individuals. Nonspecificity. Influence of drugs and training (e.g., to dampen or obscure differences between baseline and signals). Masking, in some cases (e.g., sunglasses or plastic surgery) Some differences exist (perhaps not critical) across culture. Masking (e.g., sunglasses, plastic surgery, or Botox for facial), but this may also be an indicator.

have found the framework qualitatively useful for discussing the issues that arose in our critical survey.

Figure S.5 uses a radar/spider plot to characterize a given detection system along seven dimensions, with a score of 100 corresponding to a system that has been optimized along that dimension while considering feasibility and cost-effectiveness. The score given to a lesser system is a rough and subjective characterization of how much has been accomplished relative to what would be accomplished optimally. The dimensions relate to (1) appropriate layering, (2) sensitivity and selectivity of subsystems for information when it is obtained and countermeasures are absent, (3) behavioral stimulation, (4) countermeasure resistance of those subsystems, (5) the ability to obtain the information in desirable ways that may include automated observations from a distance, perhaps without subjects being aware of the observations, (6) information fusion, and (7) mitigating the consequences of false alarms when they occur.

These dimensions relate to overall system effectiveness as shown in Figure S.6, which highlights the need for both effective detection (minimizing "false negatives") and management of the false-alarm problem (minimizing "false positives" or mitigating their negative consequences). Returning to Figure S.5, we see that it illustrates notionally

Figure S.6
Factors Affecting Overall System Effectiveness



RAND *RR215-S.6*

what progress might look like over time. For the example, we characterize the baseline (today's system) as more advanced in some respects than others, with layering having been taken seriously for some time, but with information fusion, for example, being relatively primitive and with too little work having been done to mitigate the consequences of false alarms when they occur. Progress would correspond to systems with contours farther and farther toward the extremity of the radar/spider plot (cautions in interpreting such plots are discussed in the main text). Subsequent paragraphs touch on each dimension.

Appropriate Layering

The value of layering in detection systems is well discussed elsewhere and merits little discussion other than to note that the fatal flaw in some assessments is assuming that the various layers are independent, which is not the case when, for example, they share lax or incompetent management. How much layering is appropriate depends on many factors.

Subsystem Sensitivity and Selectivity

Screening can be based on many types of information, such as background checks, overtly observable characteristics (including the carrying of weapons), or behavioral cues. Ongoing research is a mix of laboratory- and field-based empirical research and modeling. As an example, the Department of Homeland Security's (DHS's) Screening of Passengers by Observation Techniques (SPOT) Program was designed to help behavioral detection officers (BDOs) identify persons who may pose a potential security risk at Transportation Security Administration (TSA)–regulated airports. It focuses on behaviors and appearances that deviate from an established baseline and that may be indicative (imperfectly, to be sure) of stress, fear, or deception. BDOs may refer some of those whom they observe to additional screening. DHS conducted a large randomized trial of SPOT effectiveness and reported that the program resulted in the detection of illegal behaviors at a significantly higher rate than random selections, with a small false-alarm rate. Another empirical effort, Project Hostile Intent (later renamed Future Attribute Screening Technology [FAST]), included consider-

ation of hard-to-suppress micro-expressions as it sought remote, noninvasive, automated sensor technology capable of real-time detection of deception.

Although not behaviorally oriented, a second class of screening method should be mentioned. It is illustrated by the "Trusted Traveler" concept, which screens for those who can be *excluded* from some secondary screening. A 2011 analytical review demonstrated considerable promise, especially if measures are taken to deter attacker efforts to get into the program; aspects of the concept are now operational.

A constant issue in screening is how to trade off detection rate against false-alarm rate. Doing so should depend on context. During a period of high alert, security personnel can use less discriminate behavioral (and other) cues if they have additional temporary resources. During such a period, the public also tends to be more forgiving of inconvenience and somewhat higher false-alarm rates are tolerable.

As mentioned earlier, measuring physiological responses is most effective when measuring actual brain activity rather than downstream effects such as flushing. For example, electroencephalogram (EEG) measurements have shown high effectiveness in laboratory experiments to detect deception by subjects in mock terrorist attacks. Such methods, however, require the cooperation or coercion of individuals and expensive monitoring equipment as well as credible prior intelligence about the details of a potential attack. The approach appears to be countermeasure-resistant, but particularly aggressive individuals show less of the response being monitored, which would reduce real-world effectiveness. Nonetheless, the successes are a remarkable advance. Numerous empirical studies have also been performed on "downstream" responses, including use of polygraph methods and remote detection of peripheral physiological signals. These and other methods have shortcomings, such as nonspecificity and sensitivity to the "base rate" (the fraction of observed individuals who present an actual threat). If the base rate is very low, then false alarms are high.

Behavioral Stimulation

Probing to stimulate behavioral responses can sometimes improve detection effectiveness significantly. The basic concept has long been

familiar to law enforcement and many tangible examples exist, partly as the result of U.S. security officials learning from extensive Israeli practice. More generally, probing refers to the intentional stimulating of behavioral responses, such as by verbal questioning, anxiety-raising changes of procedure or process, subliminal stimuli, or tests with polygraph or EEG equipment. Probing may be polite, unobtrusive, or "in-your-face." Some probing can definitely improve detection-system results, but related experimentation and formal assessment has not been pursued as far as it might be. Verbal provocation and human assessment of verbal and behavioral responses can be effective in some circumstances without the use of sophisticated or expensive biological monitoring equipment. Israeli airport and other transit officials have used such techniques for many years, apparently with success (some of it deterrence-related). Subjective assessment of the plausibility of reasons given for traveling, or being at a certain location, along with the consistency of stories over time together provide the best clues about hostile or deceptive intent. Further research should address contextually distinct tradeoffs between benefits for detection effectiveness and negative consequences for civil liberties, commerce, and the perceived legitimacy of the security system.

Allowing for and Dealing with Countermeasures

Much of the literature and even more of the advocacy-related discussion focuses on detecting behavioral responses in the absence of countermeasures, but countermeasures are in fact a big problem, and vulnerability to countermeasures should be a prime consideration in evaluating investment programs. That said, countermeasures often are not employed, are attempted poorly, or themselves create indicators. Thus, a balance must be struck in analysis: Worst-casing could eliminate valuable measures, but optimistic assumptions could waste resources and divert attention from more promising methods. Unfortunately, net judgments are often made informally and ad hoc. Analysis could improve this situation.

Observation Circumstances: Remoteness, Covertness, Automaticity

Many of the potentially attractive technologies and methods currently depend on such relatively benign circumstances as close-up observation by humans, sometimes with a need to minimize "noise" relevant to detection systems. Operational value, however, will be much enhanced by improved capabilities to make observations from a distance, automatically, and in some instances without the subjects being aware of the observation. Progress is being made by active technology efforts on all of these. Some of the efforts are benefiting from commercial and law-enforcement-system investments in, e.g., ubiquitous security-video recordings; supervised and unsupervised computer search of data, including "big data"; and new analysis techniques, such as those used in data mining.

The Potentially Critical Role of Information Fusion

We found nothing on the horizon that presented a "magic bullet" for threat detection, raising the potential importance of effective information fusion. We reviewed quite a number of methods for combining information, ranging from very simple to more sophisticated methods. Notably, some classes of fusion have long been used. Indeed, polygraph testing combines information from several types of physiological signal. However, we have in mind information fusion that also combines information across activities and phases. We considered a number of possibilities.

Heuristic and Simple-Model Methods include checklists and risk indexes, which are especially suitable for on-the-scene security personnel. Checklists are common already and can be of two kinds, which are sometimes referred to as positive and negative (but with different authors reversing which is which). As examples, any indicator, if met, might trigger additional screening; alternatively, if all indicators are met, secondary screening might be minimized. Index methods (scoring methods) typically characterize a risk level by summing indicator scores, or by computing a risk as the product of a likelihood and a consequence, with a score exceeding a threshold triggering additional screening. Significantly, good scoring methods often need to be nonlinear and should be empirically validated rather than ad hoc. We

also consider more complex "simple methods," such as scorecards and conditional-indicator sets.

More sophisticated integration methods are likely necessary in future information-fusion centers, which would try to incorporate behavioral indicators to overcome serious signal-to-noise and false-alarm problems. Accordingly, we reviewed mathematical information-fusion methods that might be adapted and extended (these methods are discussed in more detail in Appendix D). Bayesian updating is well understood and widely applied in other domains, but its usefulness in our context is limited by its demands for many subjective estimates of conditional probabilities for which there are and will continue to be an inadequate base, and by limitations of expressiveness. Some newer methods are based on Dempster-Shaefer belief-functions, which distinguish between having evidence *for* a proposition (such as the malign intent of someone observed) and having contrary evidence (i.e., of innocence). Evidence for both can be high, whereas if the language used were that of simple probabilities, a high probability of malign intent would imply a low probability of innocence. Dempster-Shaefer theory requires fewer subjective inputs. Ultimately, however, there are several major shortcomings in using that approach as well.

A much newer approach, called Dezert-Smarandache (DSmT) theory, has not yet been widely discussed and applied, but something along its lines has promise because it deals specifically with combining evidence from sources and sensors that produce imprecise, fuzzy, paradoxical and highly conflicting reports—precisely the type of reports encountered. For example, it allows characterizing the evidence that both A and B are true; that one or the other of A or B is true (but not both); or the evidence that A is true and the evidence that A is not true. We also reviewed, briefly, the relevance of "possibility theory," various multi-attribute theories, "mutual information" (which builds on the concept of information entropy), and Kalman filtering. The best method(s) for this problem area are not yet certain, but our review may help to generate fruitful research in this critical area.

Mitigating Costs of False Alarms

As mentioned repeatedly, a major challenge in detection systems is the tradeoff between false negatives (failure to detect) and false positives (false alarms), known as Type I and Type II errors. An understudied problem amenable to research is how the broadly construed cost of the latter can be reduced—not just by reducing the false-alarm rate, but also by mitigating such bad consequences of false alarms as wasting people's time, raising their fears, insulting their dignity, or invading their privacy. We identify three classes of initiative: (1) improve system effectiveness (a "no-brainer"); (2) reduce effects on dignity and perceived violations of civil liberties (e.g., by transparency, explanation, fairness, apology, and compensation); and (3) deter abuse by those within the security system. Progress on the latter two is highly desirable for broad societal reasons and has many precedents in law enforcement. The negative consequences of false alarms alienate people, who are then less likely to cooperate, volunteer suspicions, and support the security system.

A Core Issue in the Use of Behavioral Indicators

Many of the subjects reviewed in our study are extremely contentious. Some of the controversy is scientific, relating to whether various detection methods are scientifically sound (or, as some would have it, pseudo-science). The issue is not straightforward, because detecting attacks by subjects such as terrorists involves looking for weak signals amidst a great deal of noise in circumstances in which the "base rate" is extremely low. The consequences of detection failure are very high, but there are also profound negative consequences related to false alarms, as mentioned above.

We could not resolve the controversies in this study, but Table S.4 makes distinctions useful in discussion. It compares how various methods that use behavioral indicators can be used. All of them have deterrent or cost-imposition value (second column). Would-be attackers often fear the technology and methods and behave accordingly. All of the methods can, when properly used and in proper circumstance, be

Table S.4
Some Comparisons of Where Behavioral Methods Have Value

Method	Deterrence or Cost Imposition	Flagging for Further Routine Screening		Flagging with Prejudice for Extended Checking and Detention	Tool in Interrogation	Basis for Arrest or Conviction
		Automatic	Human			
Polygraph	Yes	No	Yes	Maybe	Yes, but	No
Voice stress analysis	Yes	Yes	Yes	No	Yes, but	No
Facial expression	Yes	Technology not well developed	Yes	No	Yes, but	No
EEG	Yes	Technology not developed	Yes	Maybe	Yes, but	No
Text or speech content	Yes	Maybe	Yes	Maybe	Yes, but	Maybe
Gait analysis	Yes	Yes	Yes	Maybe	No	No

useful in providing incremental evidence on which subjects merit closer scrutiny (third and fourth columns), although there are big variations in whether they can be used automatically, remotely, and covertly. All of the methods, if well used, can *sometimes* (fifth column) justify treating an individual with considerable concern, with subsequent assessment done "with prejudice" in the sense of being potentially extended and including detention and aggressive questioning. That "sometimes" should be understood as "occasionally," however, and the methods typically have high false-alarm rates. The sixth columns uses "Yes, but . . ." to indicate that yes, if a subject merits in-depth interrogation, most of the methods can—as part of a more complex process with skilled security officers—be useful in obtaining confessions or information, but, regrettably, they can also help generate false confessions. Abuse can occur. The last column is crucial: None of the methods, except possibly for analysis of textual or vocal content, are individually an adequate

basis for arrest or conviction. Indeed, they may not be an adequate basis for putting prejudicial information in a widely shared database (e.g., "On such-and-such an occasion, the subject manifested facial-expression behaviors correlated with posing a security risk, although other factors led to his being allowed to board the aircraft").

This illustrates one of the many unresolved dilemmas. From a purely detection perspective, and assuming a process for information fusion, it would seem desirable to collect and share all kinds of fragmentary information of varied significance and credibility. However, doing so could cause serious injustices to those affected and, in many instances, would generate suspicions when none are scientifically warranted. It is instructive that, for almost a century, the FBI has maintained "raw files" on numerous subjects of observation, with important instances of those files being misused (even though it can be argued that this occurred rarely). How much more trouble would have been created if analogous raw data had been widely shared? Such issues are matters of degree, but no common agreement exists on what is and is not reasonable. As a last example motivated by current discussions in the news (as of January 2013), consider a teenager being treated for symptoms of schizophrenia. What symptoms of violent tendencies should trigger a report to authorities that would enter a sharable database, and with what balance of positive and negative consequences? Such issues are profound. We made no attempt to resolve them except that we see a major distinction between, on the one hand, using a behavioral indicator as an increment of information in a detection system seeking to identify, without further prejudice, which individuals merit more-than-usual scrutiny, and, on the other hand, using a behavioral indicator to infer probable guilt or as the basis for arrest and conviction. It is not accidental that the U.S. justice system has major constraints on how methods such as polygraph techniques can be used.

Conclusions

We found a number of important takeaways from our survey:

- Despite exaggerations found in commercial claims and the media, there is current value and unrealized potential for using behavioral indicators as *part* of a system to detect attacks. Unfortunately, analytic quantification of that potential is poorly developed.
- "Operators" are often well ahead of the science base, which is sometimes good and sometimes bad. It is very important that programs build in and sustain objective evaluation efforts, despite budgetary pressures and the tendency to see them as mere nice-to-have items. The evaluations should be subjected to objective peer review and adequate community scrutiny, although perhaps within a classified domain. The Department of Defense and the Intelligence Community have, for example, long used the federally funded research and development centers (FFRDCs), national laboratories, National Academy of Sciences, and other special panels for credible evaluations.
- Many serious problems and errors can be avoided by up-front review of procedures by experts familiar with the subtleties of detection and screening in conditions of high false-alarm rates and low base rates. Although full validation of techniques may take years (at a time when the dangers of attack are current), existing knowledge can be used to avoide many problems that are quite significant to privacy, civil liberties, travel and commerce.
- DHS and other security organizations are experimenting with proposed methods—sometimes with laudable and ambitious scientific trials that have reported encouraging conclusions (which are difficult to judge, however, without detailed access to data and methods).
- Operators, their agencies, and the scientific community have not done enough to understand how to mitigate the bad consequences of detection systems, which invariably have false-alarm problems. Much could be done.

- Information fusion is critical if behavioral indicators are to achieve their potential. Fusion should occur not just within a given method, but with heterogeneous information across activities and phases. Methods for accomplishing this are very poorly developed. This said, it remains to be seen how much can realistically be accomplished. If the indicators being fused all have very high false-alarm rates, the fused result may be more reliable but still have a high false-alarm rate. Also, success in fusion will depend on human skill in representing fuzzy, imperfect information.
- Information generation and retrieval, integration, and sense-making will tax both automated methods (e.g., including for "big data") and perfecting human-machine interactions: Machines can process vast amounts of data, but interpretation will continue to depend on human expertise and judgment. An implications is that "optimizing" should be for man-machine cooperation, not automation.
- Very little research has been done to understand how much is enough, but, subjectively, it seems that major improvements in detection are plausible with networked real-time or near-real-time integration of information. This would include further integrating (fusing) CIA and FBI information; proximate information at checkpoints and fusion-center information; and criminal, commercial, security-related, and even whole-life information. What can be accomplished is unclear, and developing a sharper understanding of payoff potential should be a priority task for objective research and analysis.
- Such steps raise profound issues of privacy and civil liberties, but the irony is that commercial organizations (and even political parties) are already far ahead in exploiting the relevant technologies and forever changing notions of privacy.
- Investment decisions about individual technologies and methods should be informed by structured portfolio-analysis approach using something like the dimensions of Figure S.6.