Data & Analytics Tools for Agile Training & Readiness Assessment

Jared Freeman	Denise Nicholson	Peter Squire	Amy Bolton
Aptima, Inc.	Soar Technology, Inc.	Office of Naval Research	Office of Naval Research
Washington, DC	Orlando, FL	Arlington, VA	Arlington, VA
freeman@aptima.com	denise.nicholson@soartech.com	peter.squire@navy.mil	amy.bolton@navy.mil

ABSTRACT

The return of American warfighters to their bases and their garrisons presents an opportunity to bolster scarce training resources and expertise with new assessment technologies. America made a similar investment in the 20th century as it shifted its intelligence budget to supplement human intelligence gathering with technologies that unobtrusively captured data concerning the activities of foreign powers. Here, we present a unifying vision of several emerging technologies that can improve military training. Following a human systems engineering approach, we first define the functional requirements of future training and readiness assessment systems, describe the architectural requirements for providing those functions, and then describe systems for the Marine Corps and Air Force that instantiate this architecture. Next we focus on two fundamental and new components of this emerging architecture: sensors that capture human performance data unobtrusively, and big data analytics that make sensor data meaningful and actionable. Finally, we identify several scientific and technical challenges encountered during the initial implementation and planned testing of these architectures.

ABOUT THE AUTHORS

Jared Freeman, Ph.D., is Chief Scientist of Aptima, Inc., a national research & development enterprise that delivers human-centered engineering solutions to DoD. Dr. Freeman's work focuses on analysis of expertise in decision-making, design of training and measurement technologies to build expertise efficiently, and computational design of organizations to employ expertise effectively. Dr. Freeman holds a Ph.D. in Human Learning and Cognition from Columbia University and a M.A. in Educational Technology from Teachers College, Columbia University. He is a contributing editor to the journal of *Human Factors*. Dr. Freeman has authored more than 120 technical publications.

Denise Nicholson, Ph.D., CMSP, is the Director of Soar Technology's new Business Area X leading an effort to explore, identify and pursue innovative applications of intelligent systems for critical and challenging problems faced by today's users. Dr. Nicholson joined SoarTech in 2014 with over 25 years of experience in a unique combination of government, academia and industry human systems research and development. Denise has a Ph.D. and M.S. in Optical Sciences from the University of Arizona, a B.S. in Electrical Computer Engineering from Clarkson University and is a Certified Modeling and Simulation Professional (CMSP). Dr. Nicholson has authored/coauthored more than 70 technical publications.

Peter Squire, Ph.D., is a Program Officer at the Office of Naval Research (ONR). Dr. Squire leads the Human Performance Training and Education (HPT&E) program area, which seeks to develop, evaluate, and deliver scientifically proven methodologies and technologies that enable the cognitive and physical superiority of Marines.

Amy Bolton, Ph.D., is a Program Officer at the Office of Naval Research (ONR). Her programs focus on enhancing individual and team decision-making and combat effectiveness through advances in research and training that improve perception, cognition, and team coordination.

Data & Analytics Tools for Agile Training & Readiness Assessment

Jared Freeman
Aptima, Inc.
Washington, DC
freeman@aptima.com

Denise Nicholson
Soar Technology, Inc.
Orlando, FL
denise.nicholson@soartech.com

Peter Squire
Office of Naval Research
Arlington, VA
peter.squire@navy.mil

Amy Bolton
Office of Naval Research
Arlington, VA
amy.bolton@navy.mil

INTRODUCTION

American warfighters are returning home to their bases and garrisons. While they are spared from large-scale deployment, they will build and maintain their readiness for the next conflict. Historically, instructors and unit leaders have assessed warfighter readiness using checklists of training tasks completed, and judgment calls about the transfer of training from exercises to battle. These methods exploit the observational powers of experts well. However, expertise is scarce; it is costly, and its impact on the conduct of training is slow and unreliable (because experts sometimes disagree). Further, expertise is no longer the only tool for the jobs at hand: training and assessing readiness.

Consider this analogy: the intelligence community successfully replaced many human operatives with data collection technologies and analytics to scale American intelligence operations. We supplemented the spies of the Cold War with sensors in space, sky, ground, and sea. Similarly, the defense community has begun to apply emerging technologies for data collection and data analysis, to increase the volume of data it collects during training, the amount it learns from those data, and the return on its investment in training and readiness assessment.

In this paper, we examine this trend from a human systems engineering perspective. We first define the functional requirements of future training and readiness assessment systems, describe the architectural requirements for providing those functions, and then describe two systems -- implemented for the Marine Corps and Air Force -- that implement that architecture. Next we focus on two novel components that are fundamental to this emerging architecture: sensors that unobtrusively capture data about human state and performance, and big data analytics that transform the voluminous sensor data into measures that drive training adaptation and assessment. Finally, we identify several challenges of this shift towards data-rich, data-driven training and readiness assessment.

FUNCTIONAL REQUIREMENTS OF AN ARCHITECTURE FOR TRAINING & ASSESSMENT

Adaptive training and readiness assessment requires a large number of measurement and management functions. Developing systems to provide these functions is an exercise in human systems engineering, an interdisciplinary method that applies systems engineering processes to develop socio-technical and social systems. Consistent with that method, we begin by enumerating the functions required to provide adaptive training and readiness assessment. To structure the presentation, we cast the functions into the familiar OODA Loop (Boyd, 1995), a cycle that comprises four phases: Observe, Orient, Decide, and Act (see Figure 1).

To **Observe** is to gather information that drives decision making and action. In training and assessment, emerging systems must capture data concerning:

- Trainee State: A record of the trainee's current physical, physiological, and neural condition state; cognitive state (or knowledge and skills); and psychological state.
- Trainee Action: A record of the trainee's observed behaviors in the training or operational context.
- Trainee Interactions: A record of the actions of trainees with their units.
- Environment State: A record of the actual conditions of training or operation in which trainees acted and interacted. (That is, it requires a record of training as delivered, not as designed).

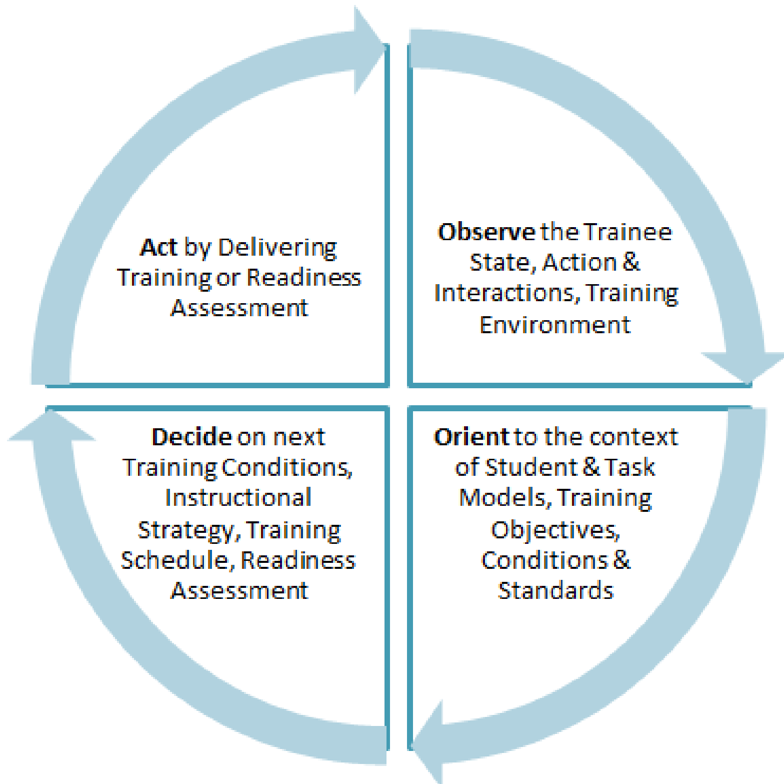

Figure 1: Essential functions of training and readiness assessment, represented within the OODA Loop.

To **Orient** is to contextualize observations in the service of making training decisions. The data that constitute context represent the current competency of the student, the competencies required of the task they train to perform, and the difference between these, which are training objectives. More specifically, the requirement for future assessment and training systems is to build, persist, and apply to observations these contextual data:

- Student Model: The history of the trainee's state, and of the training and operational events in which it was measured.
- Task Model: A description of the tasks or mission that are the subject of training and readiness decisions. (In the intelligent tutoring systems literature, this is sometimes called an expert model).
- Objectives: The goals of a training event or criteria against which readiness is assessed.
- Conditions & Standards: The criteria for satisfactory performance

To **Decide** is to choose a training action or an assessment of readiness. To make these decisions requires functions that support or make decisions concerning:

- Training Conditions: This includes training venue (e.g., classroom, simulator), the trainee's role and goal (e.g., as an MH60-R pilot, operate targeting radar to support fires), the interactions required for success (e.g., coordinate sensor and weapons operators to target a Fast Attack Craft), and constraints on these conditions (e.g., in daylight, from an altitude between X and Y).
- Instructional Strategy: How to administer briefing, training, After Action Review (AAR), and other training events.
- Training Schedule: When to administer training to build or maintain skills.
- Assess Readiness: Selecting the appropriate assessment approach (i.e., measurement instrument, collection technique), and what readiness level to assign given performance and context.

To **Act** is to administer training or to formally submit an assessment of readiness. Thus, the requirements, at the highest possible level of abstraction, are:

- Deliver Training
- Report Readiness

We are particularly concerned with adapting training to the individual and the unit. By adaptation, we mean choosing between training actions (e.g., selecting the next curriculum, course, scenario, or problem), and choosing within training events (e.g., controlling the behavior of synthetic entities, the state of the environment, or instructional events such as feedback), a distinction that VanLehn (2006) refers to as outer loop and inner loop adaptation. To adapt training in these ways requires at least the information specified above concerning the current (observed) state of the trainee, the context in which the trainee must perform, and the training and assessment actions considered in the decision.

ARCHITECTURAL REQUIREMENTS & IMPLEMENTATIONS

We are engaged in the development of two systems -- one for the Marine Corps, one for the Air Force -- that use emerging sensor and analysis technologies to satisfy a subset of the functional requirements, above. We see surprising similarities in the organization and implementation of assessment and training functions of these systems. Here, we describe that common structure, or architecture (shown in Figure 2), then turn to a brief description of each system.

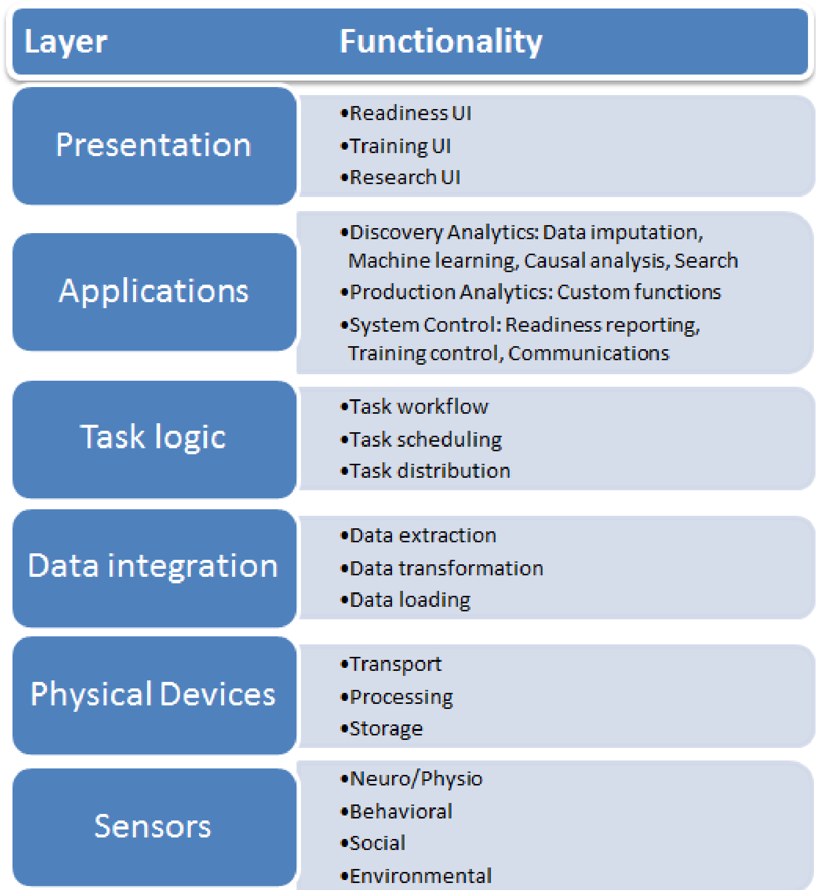

Figure 2: A layered architecture that satisfies the functional requirements for future training and readiness assessment.

This architecture consists of six layers: (a) physical sensors, (b) computer infrastructure, and (c) software for data ingest and normalization, (d) coordination of system tasks (i.e., task logic), (e) applications, and (f) presentation to users. These layers are found in other network systems for data capture, analysis, and process control. Here, they are customized both to the application -- training and assessment -- and to the scale of big data captured from neuro/physio, behavioral, social, and environmental sensors. The term "big data", as we use it here, denotes a system that may generate 1GB per second per person of data from sensors such as electroencephalography (EEG) devices, which capture electrical activity along the scalp at 1024 Hz; electrocardiography (ECG) devices, which record heart activity at 1024 Hz; eye-trackers operating at several megabytes (MB) per second; raw video computer vision analyses of facial patterns and physiological state; voice signals at several MB per minute; GPS locational data; and text chat.

The physical device layer must be optimized to store, transport, and transform data at this scale and rate. The physical device layer must include storage and high bandwidth transport characteristic of modern big data systems, whether they are implemented as centralized or distributed systems. Further, this architecture must provide the data extraction, transformation, and loading services required to capture and stream persistent sensor data, as well as the analytic functions to transform those data into measures, recommendations, and actions. These components are commonplace in intelligence, surveillance, and reconnaissance systems; medical imaging; telecommunications; and internet-scale search.

Below, we describe two implementations of this architecture: PercepTS and HUMAN. We then survey the two components that distinguish these and other emerging systems: advanced sensors of human and environmental state, and big data analytics.

Case Study: The PercepTS Architecture

The Office of Naval Research (ONR) has been exploring the use of adaptive training technologies in a methodology that follows the OODA loop process of Figure 1 for training sensemaking in the Perceptual Training Systems and Tools (PercepTS) research program (Schatz & Nicholson, 2012). Figure 3 below shows an example implementation of a layered architecture for training assessment within the PercepTS prototype system designed to work in conjunction with the Marine Corps Deployable Virtual Training System (DVTE) (Nicholson, et al., 2014). The modules that correspond to each layer of Figure 2 are as follows: a) The *Sensors* layer includes eye, head motion, and peripheral device tracking. They hook to b) *Physical Devices* for transport and computing systems that provide c) *Data Integration* in data polling and common data format layers. d) *Task Logic* is performed in view culling and visibility query layers. e) The Dynamic Tailoring System is the *Application* that controls the training experience. f) *Presentation* UIs are displayed in the Virtual Battlespace (VBS) rendering engine and supplemental Instructor/Trainee displays. As of the submission of this paper, the prototype system components have been developed in a laboratory environment and are being assembled into an integrated system. Demonstration and testing results from a field setting will be provided in the conference presentation.

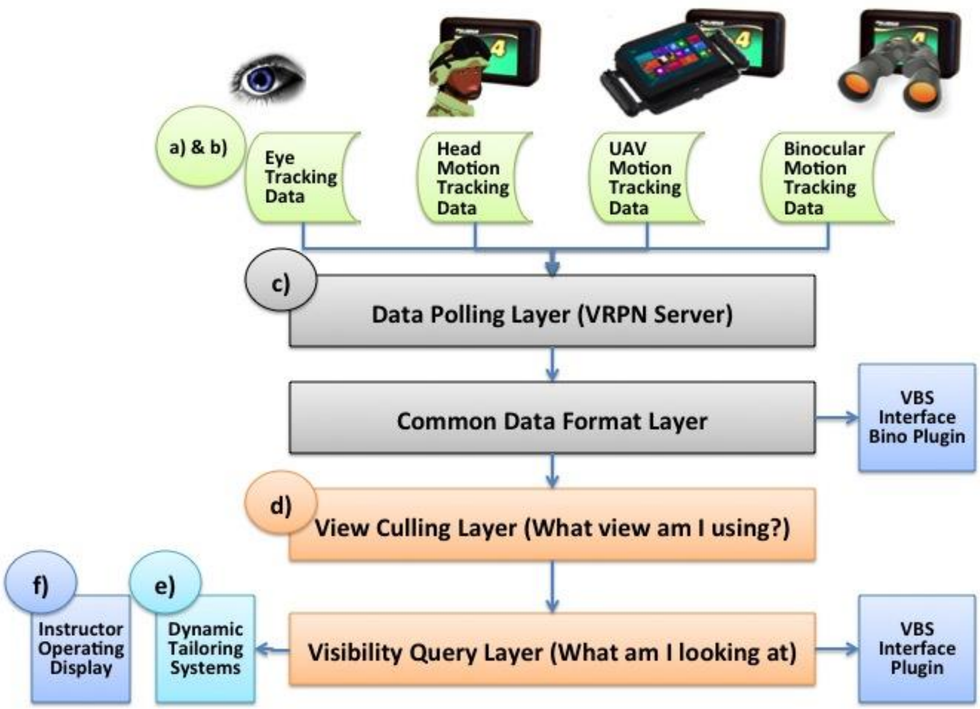

Figure 3. Example: PercepTS Data Fusion Approach

Case Study: The HUMAN Architecture

The architectural concept in Figure 2, above, is largely implemented and running daily in the Human Universal Measurement and Assessment Network (HUMAN) laboratory, operated by the Air Force Research Laboratory Warfighter Interface Division (AFRL/RHC) at Wright-Patterson Air Force Base. HUMAN is designed to prove the utility of neuroergonomic technologies for sensing, assessing, and augmenting human performance in Unmanned Air Vehicle (UAV) missions and cyber operations (see Figure 4). The systems in that laboratory implement the architecture thusly: a) The sensor layer provides technology for capturing neurocognitive signals (EEG), eye & head movement, respiration, heart activity (EKG), galvanic skin response, user actions on the keyboard & mouse, and simulation data. b) The physical layer provides a conventional, closed (not externally networked) client-server network with gigabit Ethernet connections for high throughput and large storage capacity. c) The data integration layer consists of commercial and custom sensor signal processors and Extensible Stylesheet Language Transformations (XSLT) for data transformation into a standardized Human Performance Markup Language (HPML; Stacy, et al., 2006) format, as well as a Java Messaging Service (using Apache ActiveMQ broker) and Apache Camel to enable subscription access and transport of processed data. d) The task logic layer (which is planned, but not yet implemented) will augment UAV operations and cyber mission execution by orchestrating the provision of data to applications and applications to users. e) The application layer classifies operator workload and performance in real time. f) The presentation layer visualizes this workload and underlying physiological data to support research and, eventually, supervisory action.

Two types of technologies enable these systems and other, future systems that apply the architecture we have defined above. Sensors of human state, human behavior, and context (or environment) feed the measures with which training adapts and readiness is assessed. Big data analytics enable researchers to define actionable measures from those data sources, and training systems to compute them efficiently. In the next sections, we survey these sensors and analytics.

Figure 4: AFRL's HUMAN laboratory applies the architecture, above to implement a testbed for sensing, assessing, and augmenting human performance in UAV and cybersecurity tasks.

SENSORS

Two measurement instruments currently dominate military training and assessment: written tests and human observers. The Armed Services Vocational Aptitude Battery (ASVAB) and less formal written tests are relatively inexpensive to administer and moderately predictive of performance in military career fields. Instructor and leader assessments of warfighter performance in training exercises and operations are more of a mixed bag. Experts bring judgment to the task of assessing overall competency or readiness on the basis of a few samples of human performance. They are, for example, reportedly capable of assessing the competency of an operations center by listening to the flow and tenor of communications. However, expert observers are relatively expensive, and they are not always reliable and effective. The costs include not only salary and travel costs, but the time needed to develop expertise. Experts sometimes disagree in their assessments, but the level of inter-rater reliability is typically unknown in any given training (c.f., MacMillan, Entin, Morely, & Bennett, 2013). The scarcity of experts limits the number of interactions they have with trainees, and this lowers training effectiveness. Finally, experts rarely can influence the course of training reliably in real time; thus they cannot reliably achieve adaptive training.

A third class of measurement instruments has arisen in recent years: sensors capable of capturing data concerning the state and behaviors of the individual and the collective, and sensors of the environment (or context) in which warfighters act. We carry some of these instruments with us routinely in the form of accelerometers, orientation sensors, GPS sensors, cameras, audio sensors, and heart rate monitors that reside in our smartphones, smartwatches, and exercise gear. Others sensors increasingly are embedded in our environment, such as intelligent, networked lights, locks, and thermostats that populate the growing "Internet of Things." All of these sensors are relatively inexpensive, as evidenced by the low cost of the newest smartphones (e.g., the $99 cell phone sporting the Firefox Operating System). In recent research, data from these sensors feed a variety of new measures that effectively estimate state, proficiency, or future performance of warfighters in context. Blackhurst, Gresham, and Stone (2012) envisioned the use of such sensors and measures to sense, assess, and augment human performance of the "quantified warrior" in operations. On the home front, combining sensors of humans and artifacts is termed the "Internet of Things + Humans" (O'Reilly, 2014). The same approach can enable training adaptation and improve readiness assessment. Here, we review recent advances in measurement from the level of body to behavior to social interaction.

Neuro/Physiological Sensors and Measures

The performance of individuals is some function of their expertise at a task, the task environment, and their physical and psychological state. Neurological and physiological sensors capture data concerning physical state and support some inferences concerning psychological state (Schmorrow, Estabrooke, & Grootjen, 2009).

These devices include relatively intrusive, skull-mounted electroencephalograms (EEGs), which record the location and magnitude of brain activity involved in cognition; relatively unobtrusive, wrist or chest-worn electrocardiographic (ECG) sensors of heart rate; and absolutely unobtrusive sensors, implemented as software that processes video of the head and captures heart rate, blood oxygen level, and facial expressions (McDuff, Gontarek, & Picard, in press).

The applications of these sensor data range widely. Luu, et al. (2014) applied EEG data to develop a reliable measure of intuition, defined as incipient recognition of ambiguous visual images. As part of ongoing research and development efforts funded by the United States Air Force and Office of the Secretary of Defense, Pappada and colleagues developed and validated a technology that provides real-time, second by second, monitoring and measurement of workload using only variables derived from EEG and ECG data (Pappada, et al., 2013, 2014). The technology correctly classified workload during simulated remotely piloted aircraft operations more than 83% of the time relative to the more intrusive standard, the National Aeronautics and Space Administration (NASA) Task Load Index (TLX) survey (Hart & Staveland, 1988), and 95% accuracy relative to categorizing workload dichotomously as high or low load.

These applications offer insights into the instantaneous state of individuals, the arc of their state over time, and their response to conditions in the task environment. They do so without interrupting tasks (i.e., unobtrusively), but at a cost in the volume of data generated (e.g., 1024 Hz for EEG and ECG) to be filtered, employed in calculations of measures, and stored.

Behavioral and Environmental Sensors & Measures

Measures of performance are necessarily measures of action in context. The venues for training and for assessing readiness vary from the classroom, to online venues, virtual environments, and live exercises. The sensors for assessing behavior in these contexts are also diverse.

Classrooms typically employ written tests; Sharable Content Object Reference Model (SCORM) online instruction uses the online equivalent; virtual environments (i.e., simulations) capture the experience or environment using emerging standards such as the DoD's Experience Application Programming Interface (xAPI, http://www.adlnet.gov/tla/experience-api/) and Virtual-Reality Peripheral Network (VRPN; Taylor II et al., 2001) which represent the measurable (i.e., meaningful) actions of trainees, and capture behavior through simulation controls and less widely used sensors such as eye-tracking and capture of voice communications; live exercises employ instrumentation such as the Instrumented Tactical Engagement Simulation System (ITESS) to record warfighter location and shots.

Data from these sensors is converted to measures of performance on meaningful tasks using ad hoc methods or open standards (Stacy, et al., 2006), where those measures may range from simple counts of shots and kills to sophisticated causal sequences of measures that support diagnosis of performance failure (or success) in the long chain from target detection to identification, kill, and battle damage assessment. However, these measures, too, come at a cost in data transmission, processing, and storage, notably measures that use video and audio data (e.g., at several megabytes per minute per person).

Sociometric Sensors & Measures

Readiness is primarily an assessment of the unit, from a small fire team to the full force. The readiness of the unit is not, of course, the sum of the readiness of its members. The synchronization of movement, the efficiency of communication, the provision of leadership and backup support (Salas, Sims, & Burke, 2005) are emergent properties of the team that develop as the team matures.

Sensors of teamwork include many of those used to measure individual behavior, listed above. However, devices custom-designed to capture data concerning the collective are now available. These include socociometric badges that sense audio, movement, position, orientation, and other features of human-human (Pentland, 2010, 2014) and human-machine interaction (Jones, et al., 2012).

The intriguing work lies less in the sensors than in the data mining and classification methods developed to assess team state and predict outcomes. For example, Alex Pentland has applied data concerning voice amplitude and body motion to predict the exchange of business cards and invitations to date (Pentland, 2010), and to predict the productivity and creativity of teams (Pentland, 2014).

These sensors are already at work in military research projects. For example, data from the accelerometer on sociometric sensors was used to compute measures of the synchronization of movement of small units of infantry practicing how to conduct an ambush. The accelerometer provided data that indicated whether the wearer was sitting, standing or moving. For each 10-second window, a score from 0-10 was generated for these three categories. The study found that units led by experts transited to and from destinations in few (but long) moves, and achieved great synchrony of movement between members. Units led by novices moved more frequently and that movement was less often coordinated (Pratt & Knott, 2014).

Duchon and Patterson (2014) mined exercise communications to discover (i.e., measure) emergent leaders. The measure leveraged communications that were automatically coded transcripts for a combination of indicator phrases indicative of reasoning and uncertainty. The method identified as emergent leaders those who exhibited the most reasoning and least uncertainty, and these individuals were, in two-thirds of all cases, identical to those whom observers independently identified as emergent leaders.

These new team measurement methods should help us to escape the inadequate methods of assessing readiness by aggregating individual data, and the costly and somewhat unreliable method of applying observer judgment to the problem.

ANALYTICS

It will be particularly challenging to understand and apply big data from highly instrumented training and readiness events. The challenges lie in the quality of data, which will inevitably be incomplete and thus difficult to analyze; the natural variability in performance and training conditions, which complicates comparison of trainees to each other and to performance criteria; the size of the data, which makes search difficult during discovery and application; and the richness of the data, which can obscure meaningful, actionable, causal relationships among the incidental correlations. The analytics layer of an architecture must include components that address each of these issues.

Big data imputation methods must be provided to estimate the value of missing data, such as the location of trainees during network outages in exercises, or behavior observations missing from the training record. The quality of data imputation techniques has improved markedly in recent years from the traditional practice of applying an average calculated from a similar population in similar contexts. For example, Honaker and King (2010) imputed values for missing data in survey data using a method that produced 30% better results by accounting for complex relationships between distributions of multiple variables. These methods have not been optimized for large datasets, to our knowledge. Big data implementations of data imputation methods will be required for the proposed architecture.

Big data machine learning methods must be provided to identify structure in highly variable data concerning humans and their environment. Unsupervised methods discover statistically reliable correlations of features in data. For example, Principal Components Analysis and Gibbs sampling are being tested in research to discover the knowledge that is represented by performance scores on complex tasks (Carlin, Dumond, Freeman, & Dean, 2013), the difficulty of test items on that knowledge, and the expertise of students. Latent semantic analysis (LSA and related methods) is being used discover the terms that best summarize student characteristics from their writings. Simple clustering methods are widely used to discover groups of cases that are similar on their features. While unsupervised methods reduce the complexity of data (down to a few topics, components, clusters, or features), those reductions are only guaranteed to be statistically significant and not necessarily meaningful or actionable in any operational sense. This can pose problems of interpretation and application.

Supervised machine learning methods guarantee comprehensibility of results by taking as input some assessment (e.g., good, bad) or classification for some of the data, then computing a function that estimates that assessment or classification for all of the remaining data. This approach is as accurate as the (usually human) assessment or classification. It is realized in a great number of algorithms, such as artificial neural networks and support vector machines. Machine learning algorithms are now being implemented for large scale distributed processing environments. Notable among these are the Mahout machine learning project for scalable, distributed systems (see https://mahout.apache.org/); and the Defense Advanced Research Projects Agency (DARPA) Open Catalog, a toolkit and publication library sponsored by DARPA's Information Innovation Office (http://www.darpa.mil/OpenCatalog/). The latter also contains software for efficiently searching out learned patterns of behavior in big data (c.f., Levchuk, Roberts, & Freeman, 2012).

A particularly difficult challenge is to discover patterns of events that are not only reliable, and meaningful, but are causally related. Such causal patterns help trainees and trainers understand where, when, how, and by whom decisions were made and executed. By incorporating causal explanations into analysis of training and readiness data, it is possible to deliver feedback that is more meaningful (Jensen, Chen & Nolan, 2005), accurate, and actionable. More specifically, causal patterns support diagnosis of the root cause of performance failure (and success) and inform feedback that targets the true source of performance failure (and success). They also can support feedforward guidance to prevent failures, compensatory actions by Semi-Automated Forces (SAFs), adaptations of the performance environment to enable training to continue in the face of otherwise disastrous student or system failures, and so forth. Intriguing advances in causal analysis apply Bayesian logic to large datasets to discover chains of events whose elements are necessary, sufficient, and consistently ordered. For example, Jensen, et al. (2008) have demonstrated computational methods of automatically discovering causal dependencies and natural experiments that test them in domains as diverse as academic publishing, the movie industry, and peer-production systems.

SCIENTIFIC AND TECHNICAL CHALLENGES

We have defined an emerging architecture for training and readiness assessment, one that is fit for the new environment of ubiquitous sensors and big data, one that the services can implement and develop during this period of relative peace and intensive training. Implicit in this vision are challenges of programmatics, technology, and science.

Programmatics: The infrastructure, software, and analytic services proposed here are an investment in the future force. Their purchase will come at some cost to competing programs. Some of these are training programs that return good value now. However, they do not incorporate the potential for continuous, data-driven improvement that are engineered into solutions like those above, in which performance data deepen over time, analytics are refined, and actions to assess, train, and aid performance become progressively more effective and accurate.

Technical: To store and apply diverse data -- from the physiological, to the behavioral, social, and environmental -- requires a common language, one that conveys the content and context of training and assessment data and measures. PercepTS encountered this challenge of matching the fidelity and syntax of the metadata recorded and transmitted across its layers and components, some of them independently developed components or legacy technologies. Similar issues arose in the design of the HUMAN lab. To unify the diverse sensor data and accelerate development of measures, it adopted the Human Performance Markup Language (HPML; Wiese, et al., 2006; Stacy, et al., 2006), which formally represents methods for calculating measures and assessments of human state and behavior. DoD's Experience API (xAPI, http://www.adlnet.gov/tla/experience-api/background-and-history/), is a complementary technology for capturing data at the level of the training experience.

Scientific: We are at the front of a new wave in measurement and assessment capability that is due not so much to revolutionary sensors as it is to ubiquitous sensors. The availability of data from these sensors is a promising, but pricey prospect. Research is needed to develop statistics, machine learning techniques, and experimental methods that efficiently define and compute measures that bring value to training and assessing warfighters. A case in point is the development of a real time workload measure for the HUMAN lab (Pappada, et al., 2013, 2014). While successful, this research required significant efforts in data transformation, analysis, and machine learning to hone in on a reliable and generalizable measurement formula. As the data grow in variety and size, the tools in the DARPA Open Catalog (above) and similar repositories will be increasingly useful to accelerate measure development and computation.

CONCLUSION

This article describes a new generation of systems for training and readiness assessment, systems enabled by emerging sensors and big data analytics. We have laid out the functional requirements for agile training and readiness assessment, using the OODA loop as an organizing framework. We have defined a layered structure of technologies that satisfy these requirements, and described two instantiations of this architecture -- PercepTS and HUMAN -- that prove the concept and the value of the architecture. We have surveyed the sensors that feed these or other emerging systems, sensors of the neuro / physiological state of individuals, their actions in the environment, and the social interactions of teams. We have enumerated the types of analysis tools that will be required to develop and compute measures of human performance from these big data. Finally, we have identified several programmatic, technical, and scientific problems that present challenges in the near future of data-rich, data-driven training and readiness assessment.

ACKNOWLEDGEMENTS

This work is supported in part by contracts from the Office of Naval Research: (1) Enhancing Intuitive Decision Making Through Implicit Learning (I2 Basic Research Challenge; Contract #N00014-13-C-0160) and (2) Improvement And Measurement of Perceptual, Attentional, and Critical Thinking Skills (IMPACTS; Contract #N00014-12-G-0546 0001). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here. We thank our colleagues, who participated in a 2014 workshop to define a new assessment framework: Amy Bolton, Jason Wong, Karol Ross, Jennifer Phillips, Jennifer Vogel-Walcutt, Sae Schatz, Webb Stacy, Sterling Wiggins, and Gian Colombo.

REFERENCES

Blackhurst, J., Gresham, J., & Stone, M.O. (2012). The Quantified Warrior: How the DoD Should Lead Human Performance Augmentation. *Armed Forces Journal*, December 12, 2012.

Boyd, J. R. (1995). The Essence of Winning and Losing (slide 4). Unpublished presentation. Retrieved on May 20 2014 from http://www.danford.net/boyd/essence4.htm.

Carlin, A., Dumond, D., Freeman, J., & Dean, C. (2013). Higher Automated Learning through Principal Component Analysis and Markov Models. Proceedings of the 16th International Conference, AIED 2013, July 9-13, Memphis, TN, USA.

Duchon, A., & Patterson, E.S. (2014). Identifying emergent thought leaders. In W.G. Kennedy, N. Agarwal, and S.J. Yang (Eds.): SBP 2014, LNCS 8393, pp. 50–57, 2014

Hart, S. G. & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) *Human Mental Workload.* Amsterdam: North Holland Press.

Honaker, J., & Gary K. (2010). What to do About Missing Values in Time Series Cross-Section Data. *American Journal of Political Science*, 54, (3), pgs. 561-581. Retrieved Feb 2 2014 from http://gking.harvard.edu/files/gking/files/pr.pdf.

Jensen, D., A. Fast, B. Taylor, and M. Maier (2008). Automatic identification of quasi-experimental designs for discovering causal knowledge. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Retrieved 26 April 2014 from http://kdl.cs.umass.edu/papers/jensen-et-al-kdd2008.pdf

Jensen, R., Chen, D., & Nolan, M. (2005). Automatic causal explanation analysis for combined arms training AAR. *Proceedings of I/ITSEC 2005,* Orlando, FL.

Jones, E., Lansey, J., Kern, D., Whetzel, J., Diedrich, F., & Haass, M. (2012). Automated assessment of submarine team performance using sociometric badge technology. Presented at the 2012 Joint Undersea Warfare Technology Fall Conference, Groton, CT.

Levchuk, G., Roberts, J., & Freeman, J. (2012). Learning and detecting patterns in multi-attributed network data. *Proceedings of the AAAI Fall Symposium 2012 on Social Networks and Social Contagion*. Nov 2-4, Arlington, VA. Retrieved Feb 2 2014 from http://jaredfreeman.com/jf_pubs/Freeman_PatternLearningAndDetection_AAAI_2012.pdf

Luu, P., Geyer, A., Fidopiastis, C., Campbell, G., Wheeler, T., Cohn, J., & Tucker, D.M. (2010). Reentrant Processing in Intuitive Perception. *PLOS ONE*, March 04, 2010. Retrieved on May 20, 2014 from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0009523.

MacMillan, J., Entin, E.B., Morley, R., & Bennett, W. (2013). Measuring team performance in complex and dynamic military environments: The SPOTLITE method. *Military Psychology*, Vol 25(3), May 2013, 266-279.

Nicholson, D.M., Bartlett, K., Hoppenfeld, R., Nolan, M., & Schatz, S. (2014) A Virtual Environment for Modeling and Testing Sensemaking with Multisensor Information. *Proceedings of the SPIE Defense and Security Symposium 2014.* May 5-9, Baltimore, MD.

O'Reilly, Tim. (16 April 2014). #IoTH: The Internet of Things and Humans. In *O'Reilly Radar*. Retrieved on May 20 2014 from http://radar.oreilly.com/2014/04/ioth-the-internet-of-things-and-humans.html.

Pappada, S., Durkee, K., Geyer, A., Ortiz, A., Galster, S, & Cohn, J., (2014) Developing Neurocognitive Workload Management Technologies for Multi-tasking, *Proceedings of the 85th Annual Scientific Meeting of the Aerospace Medical Asssociation*, San Diego, CA .

Pappada, S., Geyer, A., Durkee, K., Freeman, J., & Cohn, J. (2013). Modeling Operational Workload for Adaptive Aiding In Unmanned Aerial Systems (UAS) Operations. *Proceedings of the 84th Annual Aerospace Medical Association (AsMA) Scientific Meeting*, Chicago, IL.

Pentland, Alex. (2014). Social Physics: How Good Ideas Spread -- The Lessons from a New Science. New York City, NY: Penguin Press.

Pentland, Alex. (2010). To signal is human. *American Scientist*, vol 98, pg 204-211. Retrieved on Feb 2 2014 from http://web.media.mit.edu/sandy/2010-05Pentland.pdf.

Pratt, S. & Knott, C. (2014). *TRACR-T II: Tool for Rapid Assessment of Cognitive Readiness in Teams*. Unpublished technical report to Office of Naval Research, Code 30.

Salas, E., Sims, D.E., & Burkes, C.S. (2005). Is there a big five in teamwork. *Small Group Research*, 36, 555-599.

Schatz, S., & Nicholson, D., (2012) Perceptual training for cross cultural decision making. In: Nicholson, D.M., Schmorrow, D.D. (eds.) *Advances in Design for Cross- Cultural Activities* , pp. 3–12. CRC Press, San Francisco

Schmorrow, D., Estabrooke, I.V., & Grootjen, M. (2009). Neuroergonomics and Operational Neuroscience. *Proceedings of the 5th International Conference, Foundations of Augmented Cognition 2009.* San Diego, CA, USA, July 19–24, 2009.

Stacy, W., Ayers, J., Freeman, J., Haimson, C. (2006). Representing Human Performance with Human Performance Measurement Language. *Proceedings of the Software Interoperability Workshop 2006*, Las Vegas, NV.

Taylor II, R. M., Hudson, T. C., Seeger, A., Weber, H., Juliano, J., & Helser, A. T. (2001). VRPN: a device-independent, network-transparent VR peripheral system. In *Proceedings of the ACM symposium on Virtual reality software and technology* (pp. 55-61). ACM.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence and Education,* 16, 227-265.

Wiese, E., Merket, D., Stacy, W. Nelson-Walwanis, M., Freeman, J., & Aten, T. (2006). Enhancing distributed debriefs with performance measurement objects. *Proceedings of I/ITSEC 2006*, Orlando, FL.