Avoiding Big Data Overload in an Adaptive Training Use Case

Brent D. Fegley, Alan S. Carlin
Aptima, Inc.
Woburn, MA
bfegley@aptima.com, acarlin@aptima.com

Remco Chang
Tufts University
Medford, MA
remco@cs.tufts.edu

Mitchell J. Tindall, John P. Killilea, Beth F. Wheeler Atkinson
Naval Air Warfare Center Training Systems Division
Orlando, FL
mitchell.tindall@navy.mil, john.killilea@navy.mil, beth.atkinson@navy.mil

ABSTRACT

"Big data" (data often characterized by its volume, variety, velocity, veracity, and value) has been touted as the antidote to myopia and poor outcomes in data analysis and decision-making; yet this antidote only works if the recipient truly understands where big data succeeds and fails. A complex adaptive system (where understanding the parts does not imply understanding of the whole) is a case where big data alone is not the solution, because these vast, dynamic and constantly growing data introduce problems of explainability. Our present work concerns the comprehension of such a system, using big data and adaptive training as catalysts for a two-pronged approach to uncertainty and change over time. Our approach is separately prescriptive and descriptive. We focus on team training and how measurement of team performance may be made comparable, despite changing tactics, techniques, and procedures (TTPs). In our prescriptive approach, we specify a Probabilistic Graphical Model that inputs team performance across exercises, and we show how measures under different TTPs can be used to derive assessments of team and individual readiness and used, in turn, to prescribe training. In our descriptive approach, we extend techniques of automated machine learning (autoML) to help instructors explore and comprehend training data results at different levels of granularity (individual, crew, team-of-teams) and from different points of view (using the end-user's own inquiries to suggest other features of interest in the data). Our combined approach leverages machine learning algorithms for big data and the critical thinking skills of the human end-user.

ABOUT THE AUTHORS

Dr. Brent D. Fegley is a Senior Research Engineer at Aptima, Inc., where he is engaged in quantitative research, statistical inference, machine learning, and signal processing. Dr. Fegley received his Ph.D. in Informatics and M.S in Library and Information Science from the University of Illinois at Urbana-Champaign. Additionally, he holds an M.A. in Musicology from the University of Michigan, Ann Arbor, and a bachelor's degree in Music Theory/History from West Chester University of Pennsylvania.

Dr. Alan S. Carlin is a Principal Research Engineer at Aptima, Inc. within the Learning and Training Systems (LTS) division, where he has developed intelligent training systems, aircraft pilot alert systems, and data mining systems. Dr. Carlin received a PhD in Computer Science from the University of Massachusetts, an MS in Computer Science from Tufts University, and a dual BA in Computer Science and Psychology from Cornell University. Prior to joining Aptima, Dr. Carlin was Associate Staff at the Massachusetts Institute of Technology (MIT) Lincoln Laboratory. As part of his MS, he also completed the MIT Lincoln Scholar's Program.

Dr. Remco Chang is an Associate Professor of Computer Science at Tufts University. Dr. Chang leads research into big data visual analytics and machine learning affecting interactive visual analysis. Prior to joining the faculty at Tufts, Dr. Chang served as a software engineer for Boeing Corp. Dr. Chang has a PhD in Computer Science from the University of North Carolina Charlotte.

Dr. Mitchell J. Tindall is a Research Psychologist at NAWCTSD in the BATTLE Laboratory. He works in several areas such as HCI, data management and analytics, training systems enhancement and validation, and systems

software evaluation. His Ph.D. is in Industrial-Organization (I-O) Psychology from University of Central Florida (UCF).

Dr. John P. Killilea is a Research Psychologist supporting the Naval Air Warfare Center Training Systems Division (NAWCTSD) in the Basic & Applied Training & Technologies for Learning & Evaluation (BATTLE) Laboratory. He holds a Masters in Modeling & Simulation, and a Ph.D. in the same field at UCF.

Ms Beth F. Wheeler Atkinson is a Senior Research Psychologist at Naval Air Warfare Center Training Systems Division (NAWCTSD), and lead of the Basic & Applied Training & Technologies for Learning & Evaluation (BATTLE) Laboratory. She manages several research and development efforts devoted to investigating capability enhancements for training and operational environments. She holds an M.A. in Psychology, Applied Experimental Concentration, from the University of West Florida.

Authors' Note: The views expressed herein are those of the authors and do not necessarily reflect the official position of the DoD or its components. Sponsors for the efforts discussed herein include the Small Business Innovative Research/Small Business Technology Transfer (SBIR/STTR) and Naval Air Systems Command (NAVAIR) PMA-205 Air Warfare Training Development (AWTD) program.

Avoiding Big Data Overload in an Adaptive Training Use Case

Brent D. Fegley, Alan S. Carlin
Aptima, Inc.
Woburn, MA
bfegley@aptima.com, acarlin@aptima.com

Remco Chang
Tufts University
Medford, MA
remco@cs.tufts.edu

Mitchell J. Tindall, John P. Killilea, Beth F. Wheeler Atkinson
Naval Air Warfare Center Training Systems Division
Orlando, FL
mitchell.tindall@navy.mil, john.killilea@navy.mil, beth.atkinson@navy.mil

INTRODUCTION

Throwing "big data" at a problem can be like throwing water on a grease fire---not only is the fire not contained, it is likely to spread. In this paper, we consider data related to training aircrews in the US Navy, specifically those of the Boeing P-8A Poseidon, whose readiness and qualification for deployment depend critically on proficiencies (both individual and crew) acquired over many hours of training under many different conditions. Qualification depends, in part, on the crew meeting or exceeding certain thresholds of performance on specific tasks. However, crew performance is not simply a function of the crew's capability but of the crew's capability conditioned on such things as equipment (malfunction or failure), environmental conditions, the vagaries of individual evaluators (because some evaluations are subject to interpretation), and changing tactic, techniques, and procedures (TTPs; given the adaptive nature of warfare). Understanding the effectiveness of a training regime across aircrews and over time means controlling for all of these variables, and more. While collecting all kinds of data might seem like a sensible approach to addressing this challenge, not all data are equally important, and without some scheme to put them into context, the data may mislead.

How are matters handled today; and what are the limits of current practice? The collection of immense sets of data is not a new concept for the US Navy and has been an integral part of its culture for much of its history. Training wings and squadrons were required to keep track of certain pieces of information that would, in turn, be passed up the chain-of-command to help evaluate individual, crew, squadron, wing and force-wide proficiency. Unfortunately, the amount, accuracy, quality, utility, and timeliness of data provided to command leadership was greatly limited by available technology, resources, and storage and retrieval capacities. Advances in simulation-based training technology, automation, and *cloud-based* server technology has alleviated many of the challenges associated with the collection, organization, integration, analysis, and interpretation of these *big data* sets (Atkinson, Tindall, Sheehy & Bailey, 2016). Additionally, these technologies enable almost real-time assessment of proficiency. For example, data from simulated or live training events are uploaded to servers immediately following the completion of the event and are ready for integration and analysis. While these advances in training and data storage technologies have clear benefits, they also create new problems that must be addressed if we are to get the most out of our data. For example, when you collect and integrate data in near real-time, specifically in environments that are inherently dynamic (e.g., military), you can quickly run the risk of making invalid assertions about your findings. The US Department of Defense (DoD) is still in the early stages of being good stewards of *big data* and, as a result, is still understanding *big data's* benefits and limitations. *Big data* requires context to be useful. In this paper, our context is a theory of operations about aircrew training and the corresponding need to comprehend and think critically about the data collected.

What is new in our approach; and why do we think we will be successful? Our purview is holistic; by incorporating multiple different elements at different grains into analysis, we expect to improve our understanding of our domain, which is likely to lead to insight and more informed decision-making. Implicit in our approach is the recognition that we know things about the context and provenance of our data that allow us to specify relationships among various inputs mathematically and then to make inferences about those inputs. However, we recognize that one cannot know everything about one's data, especially data representing a complex adaptive system. This is why we need to remain critical of our mathematical models---using them as tools for insight rather than proclamations of state. This is why we also need mechanisms to explore the data, to help us identify gaps in our understanding. Playing "what-if" games

with our model parameters can lead us toward such understanding, but so can exploration of raw data and descriptive statistics.

Who cares? What difference will our approach make if we are successful? The desire for systems and technology to assist in the management of *big data* within the DoD is well-documented. Vice Admiral Dunaway (2015) implemented a Naval Air Systems Command (NAVAIR) data strategy requiring alignment of resources to support readiness and predictive, tactical, and strategic use of data while leveraging the science of learning to optimize aircrew proficiency. Major Blair (2015), of the US Air Force, challenged the military to better utilize the vast amounts of data we already collect for training and decision-making purposes. Additionally, Rear Admiral Morley (2016) offered strategic guidance requesting open architecture systems that are modular, scalable, and interoperable across platforms. DoD leadership has moved past simply recognizing the criticality of the collection of data and clearly understands the challenges and opportunities that exist when systems are developed to manage streams of data. Our approach is first about methodology (a way of thinking), second about technology (the supporting apparatus). If successful, we expect that our advances in both will facilitate the use and utility of *big data* in ways yet unimagined.

METHODOLOGY

Our approach is simultaneously top-down (prescriptive, confirmatory) and bottom-up (descriptive, exploratory). (See Figure 1.) We theorize that comprehension of aircrew training over time depends on both components. The prescriptive portion uses an existing data model to draw conclusions about trainees from measurement data. The descriptive portion uses new training data to refine that model. These two components can inform one another.

Our top-down component is predicated on the idea that we have some knowledge about how training works in the domain of interest and that we wish to use those assumptions to drive training. If applied to aircrew performance training data, standard approaches would require data to exhibit invariant comparison (time-invariance, for example) and sample independence. For example, in higher education, scores on high-stakes tests such as the SAT and GRE are among the criteria used to filter university applicants. The success (continued use) of these tests depends on the comparability of test scores over time despite variability among test questions. This is the realm of Item Response Theory (IRT; Lord, 1980), which is often employed to verify and validate test effectiveness. IRT is a family of models that focuses on a trainee's performance on an item (a test question or task) and establishes the probability that measured performance will be acceptable, given trainee's skill level and item difficulty.

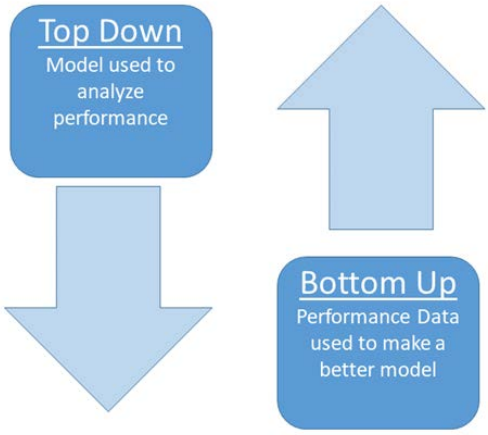

Figure 1. Concept of operations. Our Top Down component uses a model of the training domain, training exercises, metadata on training exercises, and student learning curves to analyze incoming data. Our Bottom Up component uses new data, in part, to refine the model.

However, in training, the assumptions made by classic IRT break down, in a number of ways. First, during training, trainee skill level changes as trainees improve. At the beginning of a training exercise, skill level may be low, but after the exercise, the skill level improves. Bayesian Knowledge Tracking (BKT, Corbett & Anderson, 1995) models this, and is combined with IRT in our approach. Second, Navy tasks are performed by teams of people, not just individuals, so each performance measure may be relevant to only some of the trainees, and each trainee will have his or her own skill level. In our approach, IRT is extended to account for multiple trainees. Third, in operational Navy settings, "items" are not static constructs. Rather, the Navy performs and encounters TTPs that change over time; teams that have mastered one TTP may find themselves encountering a new TTP, and their mastery may or may not be relevant to the new TTP. Thus, we need to incorporate more detail about training items in modeling. More specifically, a measure during a training exercise should not be reduced to a simple "item" identifier; rather, the descriptors (metadata) associated with that measure and exercise should be used in evaluation.

Ultimately, training measures must be turned into assessments. We regard assessments as the constructs that give meaning to measures—they are interpretations of measurements in the context of expectations. While Performance

Measures (PMs) of the kind represented in the P-8A Wing Training Manual (WTM) are vulnerable to changing tactics, assessments represent skill proficiency and therefore persist despite these changes. The relationship between PMs and assessments can be complicated and nonlinear. Fortunately, objective PMs are sometimes accompanied by observer-based assessment labels from subject matter experts (SMEs) that may yet be supplemented by analysis of "big data".

Our top-down component uses a model that accounts for all of these challenges. The model inputs training data (measures and information about training exercises, including the TTPs used). It outputs training assessments, and can also be used to make training recommendations. However, this model contains various parameters (difficulty of exercises, linkage between TTPs within an exercise and trainee skills, etc.) that affect the conclusions of the model. Two means exist for determining the correct values for these parameters. First, SMEs can work in collaboration with data scientists to assign them. Second, performance data can be analyzed to learn the parameters. This second option is one potential outcome of our bottom-up approach.

Our bottom-up component is about exploratory data analysis, where the data elements themselves are the objects of interest. Here we consider what may be learned from associated descriptive statistics, correlations, data visualizations, and facilitated mathematical modeling (automated machine learning or autoML). Thus, the bottom-up approach is about developing capabilities that allow us to learn about our data---to find peculiarities and patterns within them that help us generate or answer related questions.

The methods we describe herein are not specific to one dataset; they are applicable to the particularly challenging domain of adaptive training generally. In the next section we describe the methods that we developed as well as proofs-of-concept using synthetic data.

RESULTS

We report here on efforts to build software to address the three challenges discussed in the last section: (a) trainee skill levels are dynamic, given exposure to training itself; (b) many latent skills exist, because in our case, and in part, training is for teams, not for individuals; and (c) changing TTPs change the identity of an item (such as a task).

We began our exploration of top-down modeling techniques using Rasch measurement and analysis. A Rasch model is often regarded as a one parameter IRT model. It is prescriptive rather than descriptive: data is fit to a model rather than the reverse to help illuminate sources of variance. It also deals more effectively with Simpson's paradox (see Kievit, Frankenhuis, Waldorp, & Borsboom, 2013): rather than introducing dependencies when collapsing dimensions, the approach endeavors to separate observations to make dependencies vanish.

In practice, one could encode scores on a test as ordinal measures and then fit those measures to a one-parameter IRT model. The variable of time could be introduced into such analysis in at least one of two ways: one by using cumulative or windowed datasets for model comparison; another by reformulating the model to account for time directly (e.g., see Hung & Wang, 2012). A more sophisticated approach to skill evolution is the well-known framework called Bayesian Knowledge Tracing (BKT; Corbett & Anderson, 1995). BKT models the probability that a Knowledge, Skill, and Ability (KSA) will move from unlearned to learned after training.

To address the first two challenges, we combined IRT and BKT into a larger model called a Partially Observable Markov Decision Process (POMDP) and used it to represent team training. With a POMDP, we can assess a trainee's skill level across several skills based on performance history and then select the optimal training for that trainee based on the optimal learning path. This model allows us to associate each trainee crew member with a skill state (on each skill) and each training item (e.g., mission) along with an applicability or relevance of the item with respect to each KSA and a difficulty level. A POMDP model contains the following constructs:

- S: a finite set of states. This set is factored into individual components, so that $S = \Pi(S_k)$. Each S_k represents a team member's state in a single KSA. For each $s_k \in S_k$, $s_k \in (0, |max|)$, where s_k represents a trainee's skill level on that KSA, and max represents the maximum possible skill level. Thus, by the above description, member $s \in S$ can be described by a vector $< s_1, s_2, .. s_k >$. Many KSAs may apply to an individual trainee; many individual trainees may have the same crew KSA. To specify this relationship more clearly, we can optionally specify:
    - The set of crew members $I = \{I_1, I_2, .. I_{\{|crew|\}}\}$

  - A mapping function $\Lambda(S_i) \rightarrow U, U \subset I$ that maps KSA S_i to the crew members that S_i applies to.
  - A related mapping function $\Lambda(I_i)$ that identifies the KSAs related to crew member i.
- A: a finite set of control actions. Each $a \in A$ represents training content. Each member $a \in A$ is described as a tuple $(< d_1, .. d_k >, < app_1, .. app_k >)$ where d_i and app_i represent the difficulty and applicability of training content a with respect to KSA i.
- Z: a finite set of observations.
- $O(Z \times S \times A)$: an observation function for each action. This function is governed by IRT. However, to account for multiple skills having different applicabilities, we vectorize a 2-parameter model.

$$p(correct) = \frac{1}{1 + e^{\sum_k app_i^t (d_i^t - \theta_j^t)}}$$

  This model is easily extended to include further parameters, or into a Partial Credit Model (Masters, 1982).
- $\tau(S \times A \times S)$: a state transition function. Define an individual transition function for each S_i using applicability and difficulty and the concept of the Zone of Proximal Development (ZPD; Vygotsky 1978). The transition probability between state *s* and state *s'* given action *a* is based on the following principles:
  - The transition probability is proportional to applicability app_i.
  - The transition probability is inversely proportional to the difference in skill level between *s* and *s'* (i.e., smaller jumps in skill are more probable than large jumps).
  - The transition probability is inversely proportional to the difference in the difficulty level of the item and the current student skill level. (This enforces ZPD).
  - An equation that summarizes these three principles is below, where $s_i' > s_i$, d and s_i are always positive, ϵ is a positive constant close to zero, and $p_1$and p_2 are model parameters. (In deployed applications, these have been set to $p_1 = 2$ and $p_2 = s_{\{max\}}$, the maximum possible skill level.)

$$\tau(s_i, s_i' | a =< d_i, app_i >) \propto e^{\frac{-(p_1)(|d - s_i| + 1)(s_i' - s_i + 1)}{(p_2)(app_i) + \epsilon}}$$

- $R(S \times A)$: a reward function for each state and action.
- γ: A discount factor over future time steps.
- $b_0(S)$: An initial distribution that assigns a probability to each state, referred to as a belief state, at time zero.

For the POMDP model, the transition and observation functions depend on values for difficulty and applicability that apply to each exercise. To determine these values, we turn to Probabilistic Graphical Models (PGMs) where values of unknown variables can be inferred given variables that are known (such as those from mission profiles). Consider the PGM illustrated in Figure 2. Rather than estimate Difficulties and Applicabilities exercises directly---a task that would otherwise require collaboration among exercise authors, instructors, and data scientists upon the creation of each and every exercise---we instead construct reusable relationships between exercise mission metadata and Difficulties and Applicabilities. These relationships require specification only once; their values can be derived either by a subject matter expert (SME) or by machine learning, from performance data exposed in our bottom-up component.

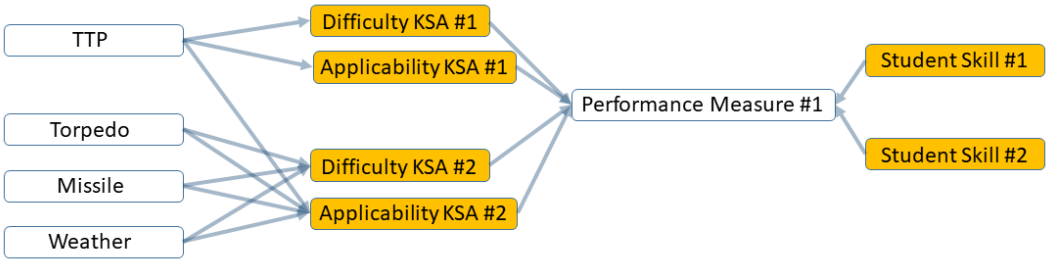

Figure 2. A notional Probabilistic Graphical Model (PGM) in which several factors influence a score on a performance measure, such as aircrew member skills and assessments of a mission's difficulty. The Tactic, Technique, and Procedure (TTP), weapons loadout, and weather determine which Knowledge, Skills, and Abilities (KSAs) are relevant as well as mission difficulty. On the right, Student Skill #1 refers to student level on KSA #1; similarly, Skill #2 refers to KSA #2.

Gold-colored variables are latent (hidden), meaning that they are not (and cannot) be observed directly; their relationships with the variables to which they connect must be inferred.

To evaluate whether such a model is effective at capturing nuances in training regime, we synthesized training exercise data based on an exercise profile template developed by our SMEs. We created 100 different teams and ran them through 20 different exercises each, selected from a corpus of 100 training exercises, each of which trains different KSAs (relevant to different team members) at different difficulty levels. Each member of each team was initialized to have a skill level of zero (meaning untrained or novice). Trainees could improve their skills after each mission. Mission success was determined by the probability of the team passing the exercise based on the difficulty and applicability of the KSA and the skill level the trainees. The probabilities associated with team member improvement and mission success corresponded to the model specified in Figure 2 and the Observation Function described above. Each of the 100 exercises in the corpus were labeled with a difficulty level from 0 to 4 (an ordinal level of measurement where 0 is least difficult, 4 is most difficult). We tested two possible training strategies: one deterministic, the other adaptive. The *control strategy* (deterministic) started at difficulty level 0, and then progressed through the difficulty levels 1, 2, 3, and 4 on the 5rd, 9th, 13th, and 17th scenarios respectively, selecting a random member from the corpus with that difficulty level. In other words, the control strategy would naively increase exercise difficulty as training continued. The *model-based strategy* (adaptive) would assemble information about applicabilities and difficulties, along with a running estimate of each team member's KSA level based on Bayesian inference, and then prescribe an exercise from the corpus based on adaptive training logic. The figure below shows the progression of average skill level of the team members over the 20 exercises on a scale from 0 to 5 (where 0 means least skill, and 5 means greatest).

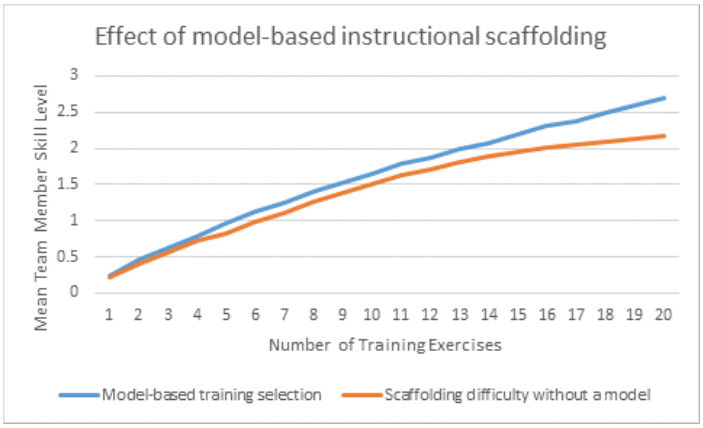

Figure 3: Comparison of regular scaffolding to an adaptive, model-based approach. Regular scaffolding advances exercises in difficulty in deterministic steps. The model-based approach advances exercises adaptively; it considers the data, in the form of each team member's performance, and then selects exercises that are best for the whole team.

As the figure shows, the model-based (adaptive training) strategy produced higher skill levels for every team member after the 20 exercises. Thus, the PGM can be used to distinguish the effects of training and highlight opportunities for improvement. By this point, the PGM has also learned (or assigned weights or probabilities to) the relationships among variables of the model. We can exploit this learning by playing "what-if" analyses---for example, testing the effect of a change in type of weather on skill level simply by varying the probability of the different types of weather we observe.

Model parameters, such as the relation between TTPs and crew skills and crew skill levels, can and should be estimated by human beings, not just automated software. To facilitate human exploration of training data, our bottom-up approach focused on "walk-up visualization", an approach to visual data analytics meant to engage an end-user in actively exploring a diverse set of data with a minimum number of interactions. One of the challenges addressed by walk-up visualization is "intent explication", a problem of human-centric computing generally. Because humans cannot express what they need to accomplish in mathematical terms, we need to infer a human intent function. Having users explore data is a means to learn that function. For example, if a user highlights an item of interest, the computer can then make inferences about what other features of the data might be of interest. This approach differs from customary exploratory data visualizations in that (a) no subject matter expertise is presumed or required and (b) sufficient time does not exist to explore the dataset with a SME's mindset. It also differs from dashboard visualizations

in that it is not designed to help a SME maintain real-time situational awareness (e.g., the health of a network or stock portfolio). The use case for walk-up visualization is simply that an end-user needs a quick overview of what might be interesting in the data.

One approach to walk-up visualization is embodied in Voder (Srinivasan, Drucker, Endert, & Stasko, 2019), an application created by Georgia Tech that presents interesting "data facts" in a tabular dataset. Generally, the application precomputes descriptive statistics about a given set of data across all its dimensions and combinations of dimensions (such as correlation, density, outliers), ranks these computed statistics by potential interestingness, and then displays the ranked list in a human-readable way (with some basic natural language processing, NLP). The ranked list may contain data facts such as the following (here, for illustration only, for a dataset about car performance):

- Most values for Horsepower are in the range 75.0 – 125.0
- Items with Cylinders:6 exhibit a moderate correlation between Acceleration and Weight
- Europe has highest average Acceleration

Horsepower, Cylinders, Acceleration, Weight, and Europe are variables in the dataset. Horsepower is treated as an interval level of measurement; Cylinders, ordinal. Figure 3 shows how these data facts appear in a prototype of the system. In the workflow, the end-user is presented with a precomputed list of data facts; the user may click on any particular fact of interest, then drill down for more detail. The system then displays a visualization of the data fact and may suggest alternative but related visualizations or data facts to the one selected by the end-user.

Figure 4. An instance of, and frames within, our prototype of the Voder user interface, illustrating data facts and select, associated visualizations. Features by number: (1) data facts, generated automatically as a pre-processing step, facilitate exploration of alternative visualizations; (2) interactive widgets facilitate end-user customization; (3) embellishments (here, a color palette) allow highlighting of a data fact within a visualization; and (4) a query interface permits search of data fact visualizations. (Source: Srinivasan, Drucker, Endert, & Stasko, 2019, Figure 2)

Voder is not the end of the story, however; our next step is to use Voder as a front-end for an automated model discovery tool called Snowcat (Cashman, et al., 2018), developed by Tufts University. (See Figure 4.) Snowcat does not require the user to have detailed knowledge of mathematical models. By design, Snowcat can handle various types of data (tabular, graph, time series, texts, image, video, audio, and speech) and problems (classification, regression, clustering, link prediction, vertex nomination, community detection, graph clustering, graph matching, time series forecasting, and collaborative filtering). However, the type of analyses available depend on the problem of interest and its supporting data types; how some of these might be accomplished, given mixed data types (e.g., multiple related tables rather than a single table) are matters of ongoing research. For example, whereas classification, regression, and clustering can accommodate any data type, generally, graph matching requires graph data, and time series forecasting requires time series data. While Snowcat can build data models of its own, its use in the present context is meant to drive understanding of the underlying data and to help explicate and improve the PGM.

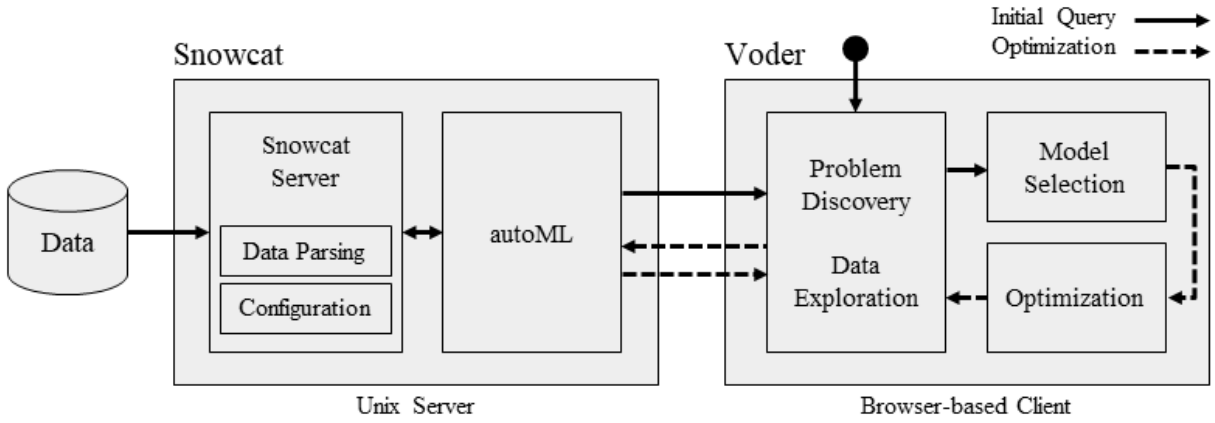

Figure 5. Conceptual architecture of a visual analytics system. Voder (Srinivasan, et al., 2019) provides a visual front-end to data ingested, analyzed, and possibly modeled by Snowcat (Cashman, et al., 2018).

DISCUSSION

What are the requirements and risks of our proposed approach? At least four emerge: the audience; model development; model update; and the forms of input. First, we have proposed two separate but complementary pieces of a puzzle, each with their own interfaces and audiences. For example, the output of the PGM (the prescriptive approach) is likely to be of interest to higher echelons in an organization, because the model itself provides a unified view in helping explicate a complex and changing landscape; insights gained from the model could lead to interventions requiring decision-making at high levels. Complementarily, the outcomes from exploratory analysis (the descriptive approach) could help explain PGM output as well as highlight idiosyncrasies otherwise marginalized out of the model. Thus, the top-down and bottom-up approaches serve two different audiences, suggesting that thought is required about the roles and needs of end-users of a system that incorporates these approaches.

Second, development of the PGM will require domain expertise---specifically, the input of those who have knowledge of, or insights about, the system under study; this expertise does not inure with casual observation. This is only the first of two challenges, however; the second is in finding the proper subdivision of the problem and the level of abstraction needed to characterize the problem mathematically. For example, takeoff and landing are critical skills for aviation, but they may not be particularly relevant for warfighting. Similarly, a low-level task such as pulling back on the throttle is not likely to be useful or tractable in a mathematical model with potentially hundreds of variables---perhaps better to concentrate on the goal of certain tasks, such as maneuvering away from heat-seeking missiles.

Third, how frequently the PGM is to be updated remains an open question. For example, a model too out-of-sync with reality may imperil decision-making; but a model updated with each observation may imperil our ability to test our assumptions (about model fit). Introducing time as a variable in the model is one way to account for change over time; creating snapshots of the PGM is another, albeit a less flexible and arguably less robust technique long-term.

Fourth, our descriptive approach currently relies on tabular data, due in part to conceptual and technical hurdles in representation. For example, the methodology for allowing an end-user to manipulate data in the form of graphs (with vertices and edges), geospatial entities, or time series is not yet clear.

How much will our approach cost? The cost of our proposed approach is largely a function of the time and effort needed to understand the target domain, and to build, test, and evaluate the PGM. Our bottom-up approach is a software development effort, in large part. However, once that software is built, it can be applied in any domain and serve to bootstrap PGM development; in other words, the software's utility will be immediate.

How long will development of our approach take? Following from the previous response, the bottom-up (descriptive) approach is likely a relatively short-order (near-term) endeavor. Once the framework is in place, the system could be used to develop intuition about the data or to inform prescriptive model-building. The top-down (prescriptive)

approach is a longer-term endeavor, because it requires access to data and thoughtful construction of relationships among model variables.

What are the mid-term and final exams to check for success? Presently, given the complexity of the system under study, the training of P-8A aircrew, a successful mid-term goal for our prescriptive approach would be the characterization and modeling of a single task for one or more aircrew members. (This requires definition of latent variables representing knowledge, skill, ability, applicability, and difficulty for the task as well as definition of the observed variables that affect performance on the task.) The final exam would be to train and then test the model with data (real or synthesized) to characterize and explain the effectiveness of the training regime that the model defines. The success of our descriptive approach will be the build-out of Voder for tabular data, secondarily for other forms of input (such as graphical and time-series data). The final hurdle will be the integration of Voder and Snowcat, which will add machine learning capability to the suite of exploratory tools.

CONCLUSION

Our holistic approach to comprehending and analyzing *big data* in a domain of interest, here adaptive training, requires the interplay of perspectives generally considered incompatible: those driven by a high-level (macro) view of the world, and those by a low-level (micro) view. While we treat these as separate components, each informs the other, and true understanding is not possible with one alone. Our top-down (prescriptive) component ensures the resilience and validity of trend analysis over time and across multiple organizations despite changes to measurement requirements. We achieve this result by separating measures and assessments and then exploiting the relationship between them. Our bottom-up (descriptive) component encourages data understanding by facilitating end-user exploration of the data itself. Highlighting features of data inputs and relationships among those inputs using descriptive statistics is a first step; automated model building, a future step. We intend that both will help in explicating surprising outcomes obtained from the top-down model, lead to improvements in that model, and promote transparency of a kind that encourages end-user trust and involvement in the process---a recognition that here, as in life itself, most things of enduring value are never fully-formed.

A grand opportunity exists to support military decision-making with data analytics. Tactically relevant aviation activity, both live and simulated, is driven by an ever-increasing corpus of revelatory data. These data, exploited by the proper analytics, can deliver inferences about the effects of changes to common, or overlapping TTPs and qualification; they can support predictions about future changes to such effects and ultimately help us to better understand force proficiency. Getting smart about (a) selecting, prioritizing, and organizing data for decision-making and (b) identifying and checking our related assumptions are prerequisites for developing such capabilities.

ACKNOWLEDGEMENTS

This research was supported by the U.S. Naval Air Systems Command (NAVAIR), Contract Number N68335-17-C-0656. We thank our partners at Aviation Systems Engineering Company (ASEC) for their subject matter expertise, which both informed and guided our research.

The views expressed herein are those of the authors and do not necessarily reflect the official position of the DoD or its components. Sponsors for the efforts discussed herein include the Small Business Innovative Research/Small Business Technology Transfer (SBIR/STTR) and Naval Air Systems Command (NAVAIR) PMA-205 Air Warfare Training Development (AWTD) program.

REFERENCES

Atkinson, B. F. W., Tindall, M., Sheehy, M., & Bailey, D. (2016). Answering the Call for Analytics within the Maritime Patrol Community. *Interservice/Industry Training, Simulation, and Education Conference.*

Blair, D. (2015). Moneyjet. *United States Naval Institute Proceedings. 141*, 68-70.

Cashman, D., Humayoun, S. R., Heimerl, F., Park, K., Das, S., Thompson, J., … Chang, R. (2018). Visual analytics for automated model discovery. ArXiv:1809.10782 [Cs]. Retrieved from http://arxiv.org/abs/1809.10782

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253–278.

Hung, L.-F., & Wang, W.-C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics*, 37(2), 231–255. https://doi.org/10.3102/1076998611402503

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology*, 4. https://doi.org/10.3389/fpsyg.2013.00513

Lord, Frederic M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New York, NY: Routledge.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika,* 47(2), 149-174.

Naval Air Systems Command. (2015). *NAVAIR Data Strategy*. [Navy Correspondence]. Retrieved from http://www.navair.navy.mil/index.cfm?fuseaction=home.download&key=BB7FC94D-87CB-4DA4AA39AFD5D3A8193F

Naval Air Systems Command. (2016). Readiness, affordability, speed will win future for Navy. *NAVAIR News*. [Navy Online Periodical]. Retrieved from http://www.navair.navy.mil

Srinivasan, A., Drucker, S. M., Endert, A., & Stasko, J. (2019). Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 672–681. https://doi.org/10.1109/TVCG.2018.2865145

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children*, 23(3), 34-41.