Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 3, 2018
A Regional Conference of the Society for Industrial and Systems Engineering

Developing a Solution to the TRADOC Analysis Center's Big Data Problem: A Big Data Opportunity

Lee Bares[1], Daniel Davis[1], Daniel Min[1], Kenneth Rau[2], and Matthew Dabkowski[1]

[1]Department of Systems Engineering, United States Military Academy, West Point, NY 10996, USA

[2]Department of Mathematical Sciences, United States Military Academy, West Point, NY 10996, USA

Corresponding author's Email: Kenneth.Rau@usma.edu

Author Note: Cadets Bares, Davis, and Min are majors in the United States Military Academy's (USMA's) Department of Systems Engineering. Cadet Rau is a major in USMA's Department of Mathematical Sciences. In May 2018 they will commission into the US Army as Second Lieutenants, serving their country in the Field Artillery and Aviation Branches.

Abstract: As data production, collection, and analytic techniques grow, emerging issues surrounding data management and storage challenge businesses and organizations around the globe. The US Army Training and Doctrine Command's Analysis Center (TRAC) is no exception. For example, among TRAC's many tasks is the evaluation of new materiel solutions for the Army, which typically necessitates the use of computer simulation models such as COMBAT XXI. These models are computationally expensive, and they generate copious amounts of data, straining TRAC's current resources and forcing difficult, suboptimal decisions regarding data retention and analysis. This paper addresses this issue directly by developing "big data" solutions for TRAC and evaluating them using its organizational values. Framed in the context of a use case that prescribes system requirements, we leverage Monte Carlo simulation to account for inherent uncertainty and, ultimately, focus TRAC on several high potential alternatives.

Keywords: big data, TRAC, Combat XXI, Requirement Analysis.

1. Introduction and Background

The mission of the US Army Training and Doctrine Command's Analysis Center (TRAC) is to "produce relevant and credible operations analysis to inform decisions" (TRADOC Analysis Center [TRAC], 2017a), and one of its primary tasks is to conduct Analysis of Alternatives (AoAs), which are the "analytical comparison of the operational effectiveness, cost, and risks of proposed materiel solutions to gaps in operational capability" (Carlucci & Zoller, 2016, p.1). Through its AoAs TRAC provides the evidence to support multi-billion dollar decisions about what systems the Army should acquire, and operational scenarios, simulation, and statistical analysis are among its primary analytical tools. Data is its currency.

Data, or more accurately *big data*, has drawn significant attention from almost every field in the past decade. A relative term, big data refers to an amount of data which cannot be captured, stored, managed, processed, and analyzed by typical database software or hardware programs (Arthur, 2013, p.1). To put this in context, in the early 1980's, the Commodore 64, the "best-selling single computer model of all time," was released to the public (Griggs, 2011). As per its name, it had 64 KBs of internal, non-removable storage (i.e., RAM); today, mobile phones hold an average of 64 GBs. In roughly four decades, the data storage capacity of personal computing devices has increased nearly a million-fold, and the exponential proliferation of data is largely to blame. Quite simply, data is getting big.

TRAC encounters big data through the use of Combat XXI – a stochastic, high resolution simulation software, which represents land and amphibious warfare from the individual solider to the brigade combat team level (*Combat XXI*). According to simulation experts at TRAC, an average Combat XXI iteration generates up to 300 MB of data, which quickly stresses TRAC's current ability to archive results when 30 or more iterations are often necessary to achieve statistically significant results (TRAC, 2017b). Moreover, when running Combat XXI simulations, there are vast quantities of information and data that TRAC is currently unable to store and analyze. Given the ability of data analytics to reveal hidden insights, TRAC believes key findings about previously unexplored options, methods, and equipment may be lost.

With this in mind, TRAC is currently working on a big data initiative. Specifically, within five years TRAC wants to be capable of systematically looking through all the data produced by its simulations, allowing it to repurpose and utilize the results of past experiments for current or future work. As such, TRAC is looking at alternate ways to store, manage, and analyze data. Among the many exciting possibilities of this enhanced capability is the ability to perform large scale design of experiments. To this end, a use case was developed that reimagines Combat XXI as an exploratory tool – one which will enable

ISBN: 97819384961-2-7

TRAC to gauge what simulation parameter settings yield operationally relevant and statistically significant results, thereby prescribing what a system's requirements should be.

2. Methodology – The Systems Decision Process

To help TRAC realize its big data opportunity, several methods common in systems engineering were applied, notably the Systems Decision Process (SDP). As seen in Figure 1, the SDP contains four phases, each with three key tasks and a result that allows the user to transition to the next phase. Its cycle is continuous and can be used in several iterations to produce refined solutions that better meet stakeholder needs.

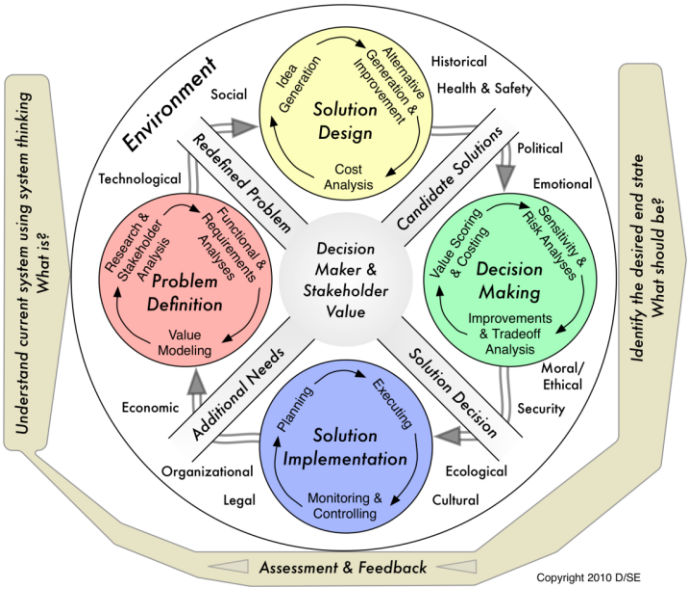


Figure 1. The Systems Decision Process (Parnell, Driscoll, & Henderson, 2011, p. 17).

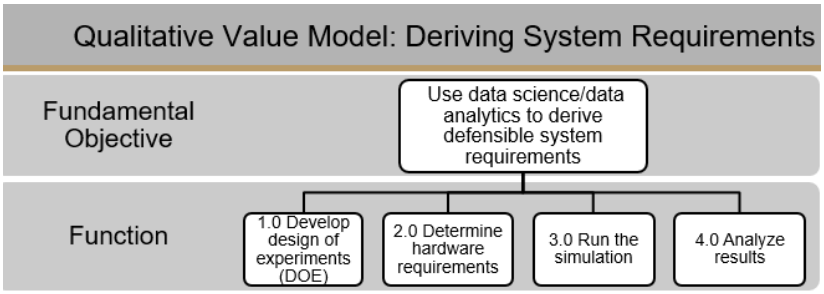
2.1 Problem Definition

The SDP's first phase, denoted in red, is *problem definition*, which includes the key tasks of research and stakeholder analysis, functional requirement analysis, and value modeling. This initial phase is critical because it frames the problem and ensures the right problem is being solved; accordingly, the end result of problem definition is an approved problem statement. With this in mind, the study team initiated the problem definition phase with stakeholder interviews. These interviews were conducted over several months, including teleconferences, video-teleconferences, and a site visit to TRAC Headquarters in Fort Leavenworth, Kansas. Many of the key insights gained through these interviews are included in Section 1, and they led directly into the subsequent functional requirement analysis. For example, the primary stakeholder (and study sponsor) COL David Tarvin, TRAC's Deputy Director, wants a system that can analyze previously gathered data in order to gain insights on current and future problems (personal communication, October 4, 2017). To enable this, the system must efficiently store large amounts of data, as well as ensure the security of classified data, due to the sensitive nature of TRAC's work. The next step was to begin value modeling by creating a qualitative value model, which has several components that paint the overall picture of the system's requirements. First is the fundamental objective, which is the highest level objective that needs to be satisfied for the system to be successful (Parnell, Driscoll, & Henderson, 2011, p. 326). Second and third are the functions and sub-functions, which capture what the system must do to accomplish the fundamental objective. Fourth are the objectives, which are always listed as maximize, minimize, or optimize some component or factor of the system. Last are the value measures, which is a scale describing how an objective is measured, always stating a factor, the units, and more is better (MIB) or less is

ISBN: 97819384961-2-7

better (LIB) (Parnell, Driscoll, & Henderson, 2011, p. 327). The fundamental objective and top level functions of TRAC's qualitative value model are given below in Figure 2.

Figure 2. Summarized Qualitative Value Model.



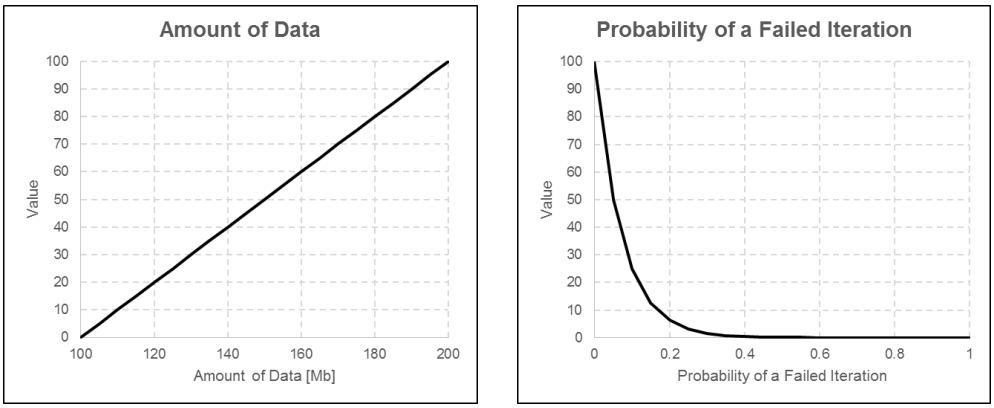
The last step in the problem definition phase is transitioning the qualitative value model into a quantitative value model, which shifts the analysis towards a more objective, technical approach. First, each value measure received an associated value function, which mathematically transforms measured performance into stakeholder value. Specifically, bounds for each value function were defined by specifying three points: what input would produce no value (minimum), what input would produce 100% value (maximum), and what input would produce 50% value (midpoint). For example, Figure 3a on the next page shows the value function for the *Amount of Data* value measure, which is defined as the amount of data saved and stored per iteration within the design of experiments. The function is linear and corresponds to a MIB objective. It ranges from 100 MB (the largest amount of data TRAC has encountered to date) to 200 MB, with a midpoint at 150 MB (50% value). Whereas the *Amount of Data* value function is rather simple, Figure 3b depicts a slightly more complicated value function for *Probability of a Failed Iteration*. This value function assesses the value associated with the chance that any one iteration of the design of experiments fails for any reason (e.g., workstation crashes due to insufficient memory, software cannot process the input, etc.). This value function is less intuitive because it is nonlinear, meaning that the returns in value are not directly proportional to the changes in the value measure. It models a LIB objective, with a maximum value at 0, a midpoint at 0.05, and a minimum value at 1. The function displays a steeply decreasing shape between 0 and 0.2 because as the probability of a failed iteration increases, failures across workstations will compound, drastically increasing the number of runs necessary to complete TRAC's simulation experiments and creating exponentially less value. Following their construction, these value functions, along with the others, were sent to the study sponsor for his adjustments, approval, and ranking.

Armed with the quantitative value model, as well as stakeholder and requirements analysis, the refined problem statement became clear, namely: *Develop a system to leverage data science/data analytics, in conjunction with TRAC's Combat XXI simulation capability, to derive defensible requirements for future materiel solutions.* To this end, a use case was developed to inspire potential solutions. Specifically, the study team envisioned a system where Combat XXI is constantly simulating new materiel systems across varied scenarios, thereby generating the data necessary for TRAC to perform prescriptive analysis of what requirements and capabilities future warfighting systems would need to increase effectiveness on the battlefield. The ultimate goal of this use case is to place TRAC at the forefront of the requirements generation process by focusing on analysis that justifies the need for new equipment.

2.2 Solution Design

In the SDP the second phase is *solution design*, which consists of three main areas of interest: idea/alternative generation, cost analysis, and alternative generation/improvement. In order to expand the design space, this phase leverages divergent thinking by encouraging the generation of broad, creative ideas for solving the decision maker's problem. Ultimately, this phase produced multiple alternatives for TRAC's future data analytics architecture.

ISBN: 97819384961-2-7



Figures 3a (left) and 3b (right): Value Functions for *Amount of Data* and *Probability of a Failed Iteration*.

2.2.1 Spanning the Design Space with Zwicky's Morphological Box

Data Storage / Transfer Hardware				Data Analysis / Simulation Hardware (i.e., Workstations)			Data Analytics Software	
Server Storage Type	Total Amount of Storage	Storage Array (RAID)	Data Transfer Options	Processor	Amount of RAM	Workstation Storage Type	General Purpose, Statistical Software	Data Visualization Software
Local All-Flash Array	5.12 PB	RAID – 6	Cable, Thunderbolt	Intel Xeon E5-2600 v3 (36 cores)	512 GB	Local All-Flash Array	Commercial JMP Pro	Commercial Tableau
Cloud	2.56 PB	RAID – 5	Wireless	Intel Core i9-7980XE (18 cores)	256 GB	Local Blended SSD / HHD	Commercial IBM SPSS	Commercial Highcharts
Local, Blended SSD / HHD	1.28 PB	RAID – 1	Cable, USB 3.0	AMD Ryzen Threadripper 1950X (16 core)	128 GB	Local SSD (Solid-State Drive)	Commercial MiniTab	Open Source Qlik Sense Desktop
Local SSD (Solid-State Drive)	640 TB	RAID – 0	Web-Based Secure Transmission	Intel Xeon E5-1600 v3 (8 cores)	64 GB		Commercial Stata	Commercial, Plotly
Local HDD (Spinning Hard Drive)	320 TB		Cable, USB 2.0	Intel Core i7-8700K (6 cores)	32 GB	Local HDD (Spinning Hard Drive)	Open Source, R	Open Source, D3.js
	160 TB		Cable, Ethernet	Intel Xeon E3-1200 v5 (4 cores)	16 GB		Open Source, Python	None

Full Monty
Middle of the Road
Bare Bones
Server Heavy
Simulation Heavy
Off Property
WCIM Informed

Figure 4: Zwicky's Morphological Box.

To generate alternatives that broadly cover (or span) the design space the study team used Zwicky's Morphological Box (Parnell, Driscoll, & Henderson, 2011, pp. 361-363). As seen in Figure 4, the columns of the box contain options for key components of the data analytics architecture, and alternatives are built by selecting a single option in each column. For example, the first alternative (*Full Monty*) selects the presumably most valuable option from each column. It will likely be the most expensive alternative, and it helps the decision maker to understand the highest performing alternative available. At the other extreme is the second alternative (*Bare Bones*), which selects the least valuable option from each column and will most

ISBN: 97819384961-2-7

likely be the cheapest alternative. The third alternative (*Middle of the Road*) seeks moderate value in each column, allowing the decision maker to understand what TRAC can have with modest improvements across all columns. The next two alternatives (*Server Heavy* and *Simulation / Analysis Heavy*) emphasize value in their respective columns and accept lower performance otherwise. The final alternative (*Off Property*) looks at remote options TRAC can utilize. It is a unique alternative that will give the decision maker a different perspective on the problem, potentially leading to new full or partial solutions for TRAC.

2.2.2 Focusing on Decision Maker Value with the Weighted Component Influence Model

Although Zwicky's Morphological Box is an effective tool for generating alternatives that span the design space, it does not explicitly consider how component options map to the decision maker's value functions and their relative importance. In order to bridge this gap the study team devised a new method – the Weighted Component Influence Model (WCIM). This method takes into account the stakeholder's interests and highlights design components that should be maximized to yield the most value. It starts by having the stakeholder rank the value functions in order of importance. In this case, the study sponsor binned the value measures into three groups (very important, important, and less important), which he subsequently rank ordered within each group. The study team subsequently used this input to assign a global importance rank to each value measure, where the most important value measure received the highest rank (17) and the least important value measure received the lowest (1). Next, these ranks were normalized, yielding *normalized importance values*.

Design components were then assessed for the extent they impacted the value measures, and they received ratings of little impact, medium impact, or high impact. As an example, the study team concluded that the *total amount of storage* has a large impact on the value measures *Number of Iterations* and *Amount of Data*. Specifically, the amount of data that is produced and collected per iteration is constrained by the total amount of storage available. Furthermore, the total number of iterations performed depends on the total amount of storage that can be filled, as more iterations generate more data, which requires more space. Similarly, the workstation's *processor* has a large impact on the *Number of Internal Factors* and *Total Wall-Clock Time per Iteration*. After all, if a more powerful processor is used, an iteration can be completed in less time, thereby decreasing the *Total Wall-Clock Time per Iteration*. Consequently, the cumulative time to execute a given number of iterations will decrease, affording more time to investigate a larger *Number of Internal Factors*. The remaining design components were assessed in a similar manner, and following these assessments, the impact ratings were assigned the arbitrary values of 0.5, 1, and 1.5 for little, medium, and high impact, respectively. This allowed the study team to take the impact values of design components, multiply them by the normalized importance values of the value measures, and sum the results across each component.

The summarized output of the WCIM methodology is a list of values that helps explain the importance and priority of each design component to the decision maker (see Figure 5). These components are subsequently emphasized in the alternative named *WCIM Informed*, which is denoted by the white arrows in Figure 4. The end state of the solution design phase is to create a list of candidate solutions which will be screened, scored, and compared in more detail within the next phase of the SDP, namely *decision making*.

Raw Influence	Weighted Influence	Zwicky's Components:
17.5	1.124	Amount of RAM
15.5	1.003	Processor
12	0.814	Storage Array (RAID)
11.5	0.745	Total Amount of Storage
12.5	0.614	General Purpose, Statistical Software
12	0.578	Data Visualization Software
8.5	0.454	Workstation Storage Type
7.5	0.448	Server Storage Type
7.5	0.408	Data Transfer Options

Figure 5: Design Component Influence and Prioritization Based on WCIM Methodology.

2.3. Decision Making

The decision making phase of the SDP consists of three main activities: value scoring and costing, sensitivity and risk analysis, and improvements and tradeoff analysis. The alternatives are screened and scored using data collected from stakeholders, research, and modeling, allowing the alternatives' performance to be quantified and ranked to see which alternative is the most beneficial. The alternatives' values are then plotted against their respective costs to show which

ISBN: 97819384961-2-7

alternatives are optimal for the decision maker. Once complete, sensitivity analysis examines how changing the weights of certain values impacts the overall scores, illuminating the role of uncertainty and allowing risk informed trades to be made. Finally, the study team will take these results and look for ways to improve the solution, ultimately recommending an alternative that best satisfies TRAC's needs.

3. Conclusion and Future Work

Through the SDP the study team has developed alternatives that will allow TRAC to capitalize on its big data opportunity. In particular, by designing a holistic big data architecture that accounts for data storage and simulation hardware, as well as analysis software, TRAC can realize its use case to leverage Combat XXI as an exploratory tool to inform defensible system requirements. In the near future, the study team will work with TRAC to find the option that best meets its needs, wants, and budget, and the goal is to have the SDP's decision making phase complete by the end of April 2018.

After TRAC makes a decision, the next step is to execute the fourth and final phase of the SDP – *solution implementation*. This phase analyzes the planning, execution, and controlling aspects of the decision, and it ensures TRAC's expectations are realized. Once the architecture has been fully implemented, a potential area for future research is how TRAC's approach can be leveraged in other analytical organizations across the Army.

4. References

Arthur, L. (2013, August 15). What is Big Data? *Forbes*. Retrieved from https://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#527d0c685c85

Carlucci, R. & Zoller, N. (2016, December). *Analysis of Alternatives (AoA) Process Improvement Study* (CAA-2016058). Fort Belvoir, VA: Center for Army Analysis. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/1029532.pdf

COMBAT XXI. Fort Leavenworth, KS: TRADOC Analysis Center. Retrieved from http://www.trac.army.mil/COMBATXXI.pdf

Griggs, B. (2011, May 9). The Commodore 64, that 80's computer icon, lives again. *CNN*. Retrieved from http://www.cnn.com/2011/TECH/gaming.gadgets/05/09/commodore.64.reborn/

Parnell, G., Driscoll, P., & Henderson, D. (Eds.). (2011). *Decision Making in Systems Engineering and Management* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

TRADOC Analysis Center. (2017, August 14). *US Army TRADOC Analysis Center (TRAC): Overview* [PowerPoint slides].

TRADOC Analysis Center. (2017, November 6). *USMA DSE Capstone Visit* [PowerPoint slides].

ISBN: 97819384961-2-7