The Cyber Data Science Process

Major General John W. Baker
Dr. Steve Henderson

Our world is facing explosive growth in data being communicated on and generated by its people, their systems, and their networks. More data has been created in the past two years than in the entire previous history of mankind (Heidorn, 2016). By 2020, our digital universe of data will grow to 44 zettabytes (or 44 trillion gigabytes) which is ten times its size today. The enormity of this data and our ability to apply advanced technology to leverage it to gain new insights is often described as the era of “big data.” The study and application of big data spawned a new interdisciplinary field known as data science which combines the domains of operations, mathematics, and computer science as well as several ancillary fields such as social science, intelligence, and economics. The application of data science has already shown great promise in a wide range of fields from medicine to business.

Because of these achievements, there is a natural expectation that the U.S. Army will equally benefit from data science, particularly in the data-rich area of cyber security. Based on data science successes in the civilian sector, the Army hopes to leverage its data to increase cyber situational awareness, maintain clairvoyance about its networks, and achieve information dominance over its adversaries. As we described in our previous article (Baker & Henderson, 2016), data produced on and by military networks defines the very contours of military cyber operations and must be mastered by the Army to gain a competitive advantage against our adversaries. In the words of Google's Eric Schmidt, “the Pentagon needs its own Google for all its data” (Defense One, 2017). In the spirit of helping the Army leverage its data at Google-like levels, we presented the case for a cadre of Army data scientists to lead this effort. Our recommendation follows the analysis of the problem and reflects trends and best practices observed in

Major General Baker is a 1985 graduate of Norwich University and was commissioned a Lieutenant in Armor. He served his initial assignment with the 1st Squadron, 3rd Armored Cavalry Regiment. He was branch transferred to the Signal Corps as a Captain.

MG Baker has commanded signal units at every echelon; company, battalion, brigade, theater, and now NETCOM, a global command.

He holds Master's Degrees from Central Michigan University and the Industrial College of the Armed Forces. He is a graduate of the Armor Officer Basic and Signal Officer Advanced Courses, Command and General Staff College, and Industrial College of the Armed Forces.

Major General Baker and his wife Laurie have two daughters, Alexis and Mackenzie.

government and non-government entities adapting to a data-fueled revolution that is impacting everything from cybersecurity to logistics to health care (The White House, 2014; Verizon RISK Team, 2015).

As the Army moves quickly to seize on opportunities presented by data, there is a natural tendency to focus on 'the what and who' aspects of a solution. What technology do we need to design, purchase, and engineer? Who do we recruit, train, and develop to use this technology? Who leads this effort? However, much less attention has been devoted to how these personnel and technologies are specifically brought to bear on cyber operations. In this paper, we outline the *Cyber Data Science Process* to addresses this question. The Cyber Data Science Process is a workflow of specific activities that define how data science should be incorporated with cyber operations. It combines the latest in data science research with doctrine and best practices found in military intelligence and targeting activities. We include a functional analysis of the workflow and identify the actions, skillsets, and products required at each stage.

Our national security requires the U.S. Department of Defense (DoD) and other agencies having guaranteed access to a reliable, secure, and accessible network at all times. This network is known as the Department of Defense Information Network (DODIN). Data science and its associated processes are key requirements to the network's security and resilience. The Army Network Enterprise Technology Command (NETCOM) provides the Army's portion of the DODIN, ensuring freedom of action in cyberspace while denying the same to adversaries. A major implied task in NETCOM's mission is gaining and maintaining complete situational awareness about what is happening on its networks. However, there are a number of challenges that make this task difficult.

Dr. Steve Henderson is a Senior Cyber Security Research Scientist working in the Software Engineering Institute at Carnegie Mellon University. He is an accomplished computer scientist and systems engineer with over 23 years of experience in the Department of Defense community defining requirements, designing solutions, implementing systems, and leading teams to solve complex technical challenges. Steve is also a retired U.S. Army Lieutenant Colonel. He holds a Ph.D. in Computer Science from Columbia University, an M.S. in Systems Engineering from the University of Arizona, and a B.S. in Computer Science from the United States Military Academy.

The first challenge is related to the evolving implementation of DoD cyberspace doctrine. Cyber operations are defined as “the employment of cyberspace capabilities where the primary purpose is to achieve objectives in or through cyberspace” (JP 3-12). Within the DoD, cyber operations fall under the purview of the U.S. Cyber Command (USCYBERCOM) and its component commands: Army Cyber Command (ARCYBER), Fleet Cyber Command, Air Forces Cyber (AFCYBER), and Marine Corps Forces Cyberspace Command (MARFORCYBER). The capable men and women of these commands are trained and equipped to handle a broad range of offensive and defensive cyber operations and work with a number of non-military agencies to secure our national interests in cyberspace. In support of ARCYBER, NETCOM personnel operate the DODIN and participate in defensive cyber operations (DCO) conducted on their wide-reaching enterprise. This exceptional team of soldiers and civil servants do a tremendous job of keeping our network safe, healthy, and online. Nevertheless, they are not staffed, trained, or equipped to handle cyber operations informed by data science at levels equivalent to their potential adversaries. And, we must assume these same adversaries will use data science techniques to get past our strategic defenses to compromise our lower level networks. If we want to maintain complete situational awareness and freedom of maneuver on these networks, it is imperative that we staff, train, and equip NETCOM personnel to conduct data science informed DODIN and DCO operations at a sufficient level of capability.

A second challenge facing the Army deals with analyzing data generated on its networks. These networks span 20 countries around the globe in support of Unified Land Operations, 32 major commands, over 800,000 people, and operating 1.1

million devices. Collectively, these users and their machines generate 20 terabytes of data daily. In order to maintain situational awareness, five Regional Cyber Centers (RCCs) monitor portions of this data for events such as network intrusions, service interruptions, and suspicious network flows. However, RCCs are not resourced to completely leverage all the data at their disposal. What is needed is an ability to conduct state-of-the-art, large-volume, near real-time data science akin to best practices employed by our partners in industry (Marr, Bernard, 2016; Russom, 2011). These analytics could enhance Indications & Warning (I&W) capabilities and bolster incident response and real-time targeting data shared with USCYBERCOM. The analytics could also help inform orders generation, identify and forecast advanced threat behaviors, tune sensors, prioritize systems administration activities, and guide engineering efforts. Fully leveraging all the data generated on our networks is essential to out-maneuvering our adversaries in cyberspace and ensuring freedom of maneuver.

While we clearly need to address these cyber operation and data science man, train, and equip challenges we identified a third, equally critical, challenge. We submit that the Army needs to develop and validate a process to guide how we integrate data science capabilities into cyber operations. Even if Army units have sufficient people, experience, training, and tools to conduct cyber operations complemented by data science, how would these capabilities be best employed? What is needed is a detailed doctrinal process that governs how powerful data science capabilities can complement and augment our current military staff and decision-making practices. This process should be based on industry best practices, support current military doctrine, and provide sufficient detail to guide how we task-organize and operate to fully leverage data science in cyber operations.

The application of data science has already shown great promise in a wide range of fields from medicine to business.

ANALYZING INTELLIGENCE & TARGETING PROCESSES

Toward this end, we looked for inspiration from two types of processes found in military science: intelligence collection & targeting. Based on our experience, we believe these two analogs offer great insight for applying data science to cyber operations. Both intelligence gathering and targeting place the enemy at the center of our analysis and complement terrain-based approaches that focus on technical infrastructure. The added emphasis on the enemy helps augment traditional security models focused on incident handling and compliance (Security for Business Innovation Council, 2012). This is especially important in cyber operations where the enemy is comprised of hundreds of attackers daily ranging from non-nation to nation-state actors, to organized criminals, to hactivists, to novice script kiddies. Most of these entities operate in isolation and are pursuing different, uncoordinated objectives while employing different tactics, techniques, and procedures (TTPs). Thinking about cyber defense as a one-size-fits-all model treats these threats equally and fails to address nuances that are exploited by the attacks. Instead, we need an intelligence-focused approach that focuses our defensive posture and can be applied across a wide array of bad actors simultaneously, targeting and out-maneuvering each with synchronized and well-coordinated cyber operations.

We examined several well-known processes from the intelligence and targeting realms. Our goal is not to replace these processes because each plays an important and established role in military operations. These processes provide support to military decision making and can be directly applied to cyber operations. Rather, our goal is to analyze the processes to determine how they can inform a more detailed and low-level process to help the Army data scientist.

The first process we examined is defined in DoD Joint Publication 2.0 (JP 2-0, 2013) which describes a general doctrinal intelligence process practiced within the DoD. This process, shown in Figure 1, is followed by each component service, though some services may choose to augment certain steps.

THE INTELLIGENCE PROCESS

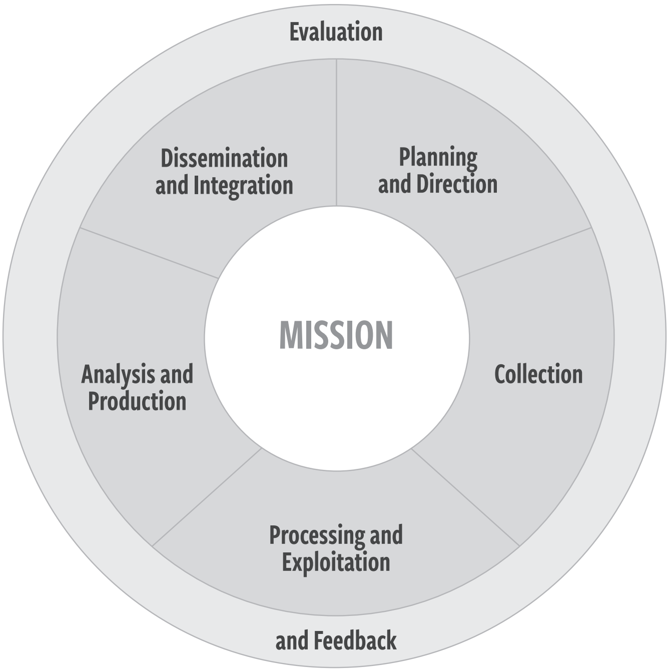


Figure 1. The JP 2.0 Intelligence Process

The DoD joint intelligence process involves five sequential phases that are centered on supporting a particular mission and reinforced by continuous evaluation and feedback. The first phase, Planning and Direction, consist of outlining the specific intelligence activities and actions required to support the mission. This includes prioritizing and directing intelligence collection efforts and assets. The collection phase of the process involves the physical act of acquiring intelligence data and information from human, imagery, signal and other intelligence sources. The Processing and Exploitation Phase involves activities to collate, clean, store, and organize collected intelligence information for follow-on exploitation and analysis. The exploitation portion of this stage involves an initial and rapid review of processed information to identify high-value and time-sensitive information that can immediately support the mission. The Analysis & Production stage is a deliberate activity to carefully study, review, and combine the various intelligence information and produce one or more intelligence products. These products include, but are not limited to, reports, estimates, briefings, and diagrams. The final stage, Dissemination and Integration, involves distributing various intelligence products to units and individuals and integrating analysis and recommendations into current and future operations.

> We submit that the Army needs to develop and validate a process to guide how we integrate data science capabilities into cyber operations.

As a potential candidate to guide data science, the Joint Intelligence Process presents several strengths and weaknesses. One strength of the model is that each of its component stages are data-driven activities that provide natural opportunities to apply data science. For example, the process and exploitation stage involves analytical tasks that are performed with data science techniques including pattern-recognition, natural language processing, and machine learning. A second strength is shown in how the continuous evaluation and feedback activity encourages a work flow where data analytics are reviewed and improved throughout the entire process. On the negative side, the process is fairly high-level and doesn't specify many details on how specific data science functions should be performed in each step. Moreover, the intelligence process isn't necessarily presented with a view toward cyber operations. Finally, the main phases of the intelligence process are sequential with no intermediate opportunities to iterate or go back to a previous step.

The next process we examined in detail is a cyber-focused intelligence process known as the Cyber Intelligence Lifecycle and developed by the Intelligence and National Security Alliance (INSA) Cyber Intelligence Task Force, Tactical Cyber Intel (INSA, 2015). The process, which is shown in Figure 2, has seven steps.

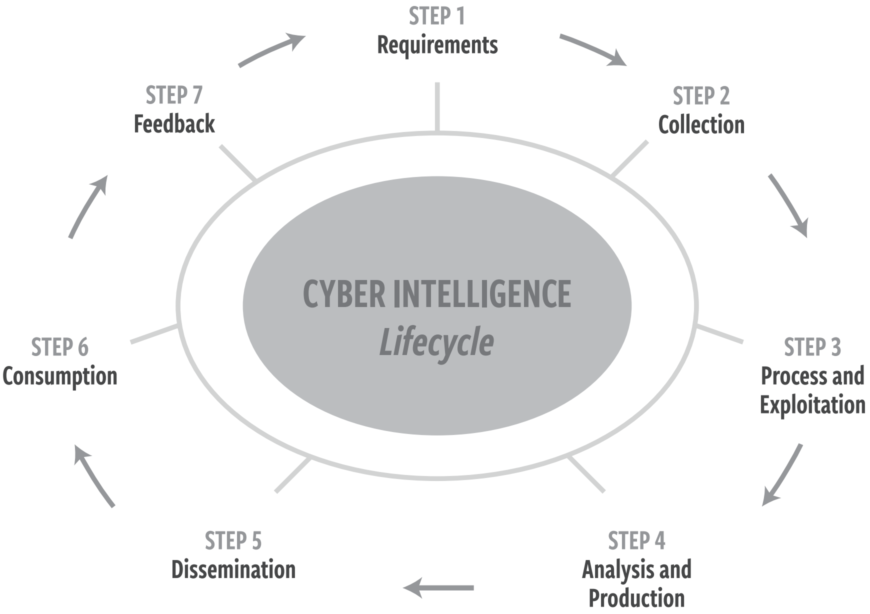


Figure 2. Cyber Intelligence Lifecycle

Step 1 defines the requirements for the overall intelligence process. This involves enumerating what intelligence products and outcomes are needed to support the mission. Requirements flow from analyzing the current environment, organizational goals, essential aspects of the mission, and previous threat intelligence. This includes deriving a detailed network map and enumerating possible data sources. Step 2–5 cover collection, processing, exploitation, analysis, production, and dissemination and are similar to related functions in the JP 2.0 joint intelligence cycle. Step 6 entails an explicit consumption function, which involves ensuring intelligence outcomes and products are integrated with the decision-making process and acted upon in a timely and sufficient manner. Step 7 entails reviewing these generated and consumed intelligence outcomes to determine if the original requirements were satisfied.

The strengths of this model are its enumeration of requirements and consumption activities as dedicated steps in the lifecycle. The requirements elicitation step ensures the process defines specific outcomes to satisfy stakeholder and mission needs. This helps keep the intelligence process agile, mission-focused, and relevant. The deliberate consumption step makes it the intelligence analyst's responsibility to ensure products they develop are consumed by the stakeholder. This encourages a continuous dialog between the analyst and the stakeholder to ensure requirements are met. The model shares the same weaknesses as the Joint Intelligence model; mainly it lacks specificity for data science tasks and it not internally iterative.

We next examined the Find, Fix, Finish, Exploit, Analysis, and Disseminate targeting process or F3EAD (U.S. Army, 2010; Faint & Harris, 2012). This process, depicted in Figure 3, is a tactical-level process developed to help Army units identify, target, and exploit high-value individuals (HVIs) across an enemy organization.

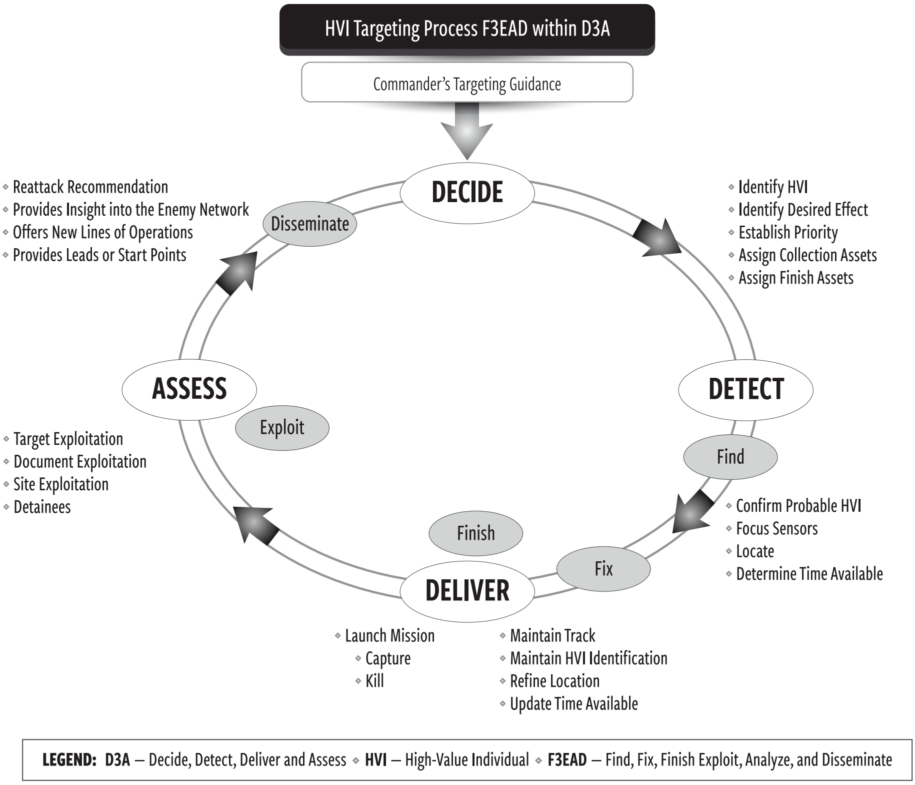


Figure 3 : F3EAD Targeting Process (U.S. Army, 2010)

The process has four high-level functions. The Decide function is the process of establishing a prioritized list of potential targets, what effect is desired for each (i.e. captured, killed, neutralized), and what intelligence and operational assets (i.e. drone, special operations, host-nation law enforcement) to apportion to each target. Detect is the process of finding and fixing each target using allocated intelligence assets. The Deliver function is launching operational assets to deliver the desired effect on each target. The delivery is followed by the deliberate exploitation of each target which includes prisoner interrogation, reviewing captured documents, and harvesting data from digital evidence. The Assess function involves analyzing this exploited information to determine new intelligence that is disseminated to inform additional operations. This includes adding new targets to the prioritized targeting list heading into the next iteration of the F3EAD process.

The strength of this model is its focus on specific enemy targets. The prioritized targeting process drives deliberate resource allocation for intelligence and operational assets. As part of a data science model, this same focus would provide explicit direction to the data science team about which analytics should be written to locate which targets. This would provide clarity and specificity to the team's efforts. However, a prioritized targeting approach would face limitations in the cyber domain. Not every threat to our networks represents a clearly identifiable entity we can track. Zero day vulnerabilities, unintended and unknown functionality caused by imperfections in the software design process, leave our systems vulnerable to the first threat actor that can identify the zero day and exploit it. While certain classes of threats would be traceable to High Value Individuals, the sheer size of our networks and the anonymity and obscurity offered by cyberspace technology make the targeting process highly dynamic and abstract. Data science can help illuminate this abstraction and may lead to refinement of how we think about targeting in cyberspace. For example, high-value targeting could be expanded to include High-Value Behavior, High-Value Organization, and High-Value Network Infrastructure. The highly iterative and agile nature of the F3EAD model can serve as an excellent framework for thinking about data-science supported targeting in cyberspace.

> As a potential candidate for integration with the intelligence and targeting models, the Data Science Workflow presents several strengths and weaknesses.

DATA SCIENCE PROCESSES

Because we are interested in the application of data science to cyber operations, we also examined data-centric processes. Relevant work traces back before the emergence of data science to the age of the database. In this era, large, single-instance, industrial-strength databases powered academic research and business operations. Great interest was placed on extracting novel information from these databases which led to the fields of Knowledge Discovery in Databases, or KDD (Klösgen, 1996; Klösgen & Zytkow, 2002), and Knowledge Discovery and Data Mining, or KDDM (Reinartz, 2002; Cios, Swiniarski, Pedrycz, & Kurgan, 2007). These fields are collectively referred to as knowledge discovery process (KDP) for which Kurgan and Musilik provide an excellent survey (2006). Notable work includes an ad-hoc model outlined by Brachman and colleagues (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, & Simoudis, 1996) which was extended to a foundational KDD model by Fayyad, Piatetsky-Shapiro, and Smyth (1996). This model defines a seven-step sequential process consisting of identifying goals, creating target data sets, data preprocessing, data transformation, data mining, pattern evaluation, and knowledge presentation. The model places particular emphasis on the data mining step which is the process of applying algorithms to find patterns in data. Another important model, known as Cross Industry Standard Process for Data Mining (CRISP-DM), was created by an industry consortia consisting of International Business Machines Statistical Package for the Social Sciences (IBM SPSS), National Cash Register Corporation (NCR), Daimler Chrysler, and the Dutch banking company Onderlinge ziektekostenverzekeringsfonds van Hoogere RijksAmbtenaren (OHRA) (Shearer, 2000). The CRISP-DM model consists of six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. It still enjoys broad acceptance in the business world.

While the KDD and CRISP-DM models provide excellent foundations for creating a knowledge management process in any organization, they are abstract models that deliberately leave specific implementation details open to interpretation. Several researchers proposed additional models to add this specificity. These include work by van der Heijden who proposed the Process Mining Project Methodology (2012). This model includes specific data science tasks such as tool selection, data preparation, and decision model validation. Sipoloa applies the Fayyad's KDD model (Fayyad et al., 1996) to identify anomalies in network traffic (Sipola, 2015). In doing so, he adds specific data mining functions such as feature extraction, normalization, dimensionality reduction, and classification to the process (Juvonen & Sipola, 2012). Guo conducted extensive research into research programming, or the process of using computer programs to obtain insights from data

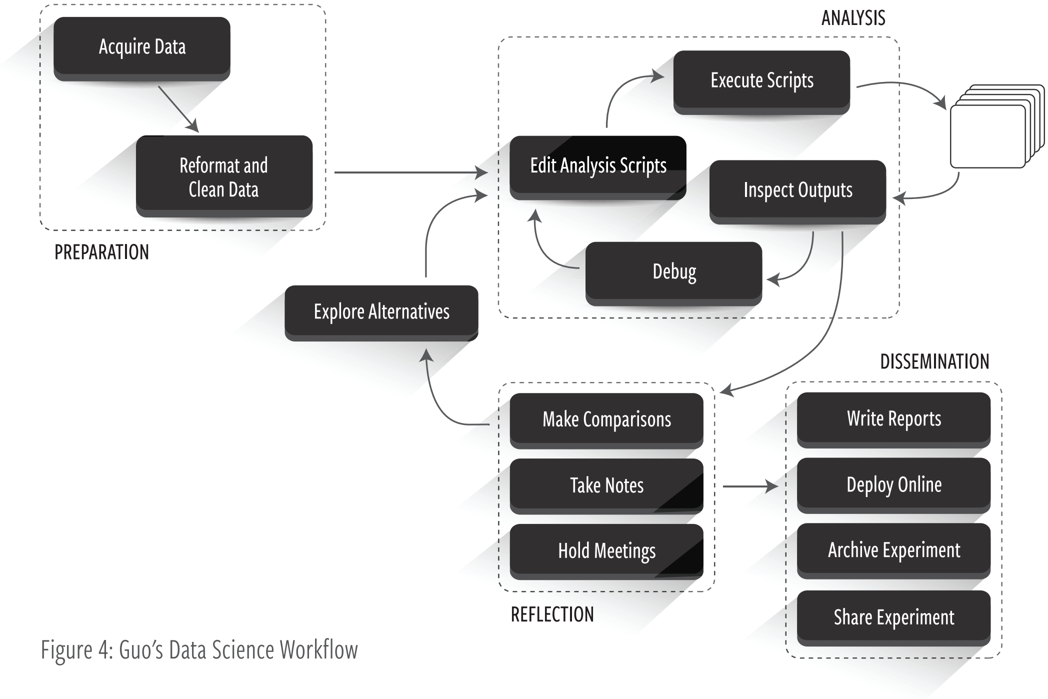


Figure 4: Guo's Data Science Workflow

(Guo, 2012). As part of this work, he proposed the Data Science Workflow consisting of four major phases (preparation, analysis, reflection, and dissemination) which each include detailed data science tasks.

Of this prior work, we selected Guo's model for further analysis because it captures the main phases found in the CRISP-DM and major KDD models while including additional detail specifically focused on data science. Within overarching stages of Preparation, Analysis, Reflection and Dissemination, Guo introduces several specific tasks. These are depicted in Figure 4.

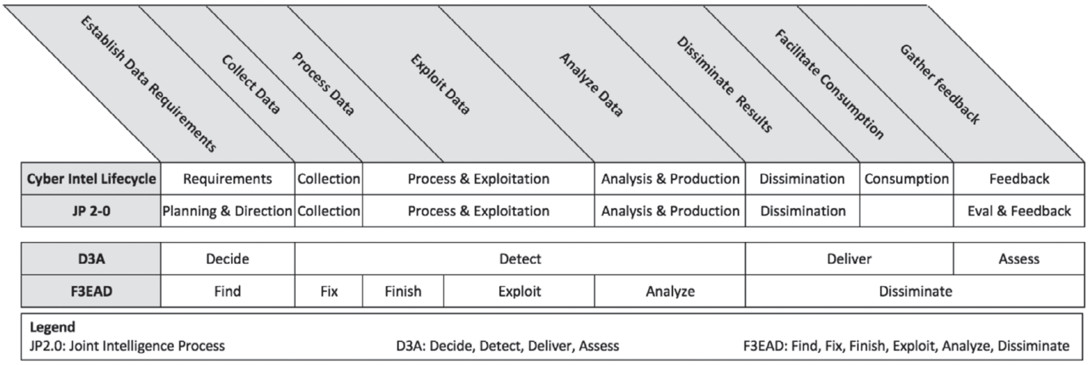
> The exploitation function involves an initial and rapid review of newly processed information to identify high-value and time-sensitive information that can immediately support the mission.

The preparation stage first involves acquiring the data, and then reformatting and cleaning it for follow-on analysis. From there, the process enters an analysis loop where the data scientist edits analysis scripts (computer programs) that are used to process the data. When these scripts are executed, they produce multiple outputs that can include statistics, tables, metrics, and charts. These outputs should provide new insights into the data scientist's questions. The data scientist inspects these outputs and, if not conclusive, debugs them, and then edits and runs them again. Once the outputs are verified, the data scientist carefully reviews them in the reflection phase. Outputs are compared against each other for accuracy and trends, and the data scientist invites others to collaborate on the findings. Documentation is critical in this phase, and the data scientist makes detailed notes about observations, limitations, and decisions made regarding the output. If required, further analytical product alternatives are explored to help confirm findings, address gaps, and eliminate inconsistencies. This spawns another cycle of the analysis loop. Once the data scientist arrives at a set of verified and validated outputs that provide new information they are finally ready for the final phase: dissemination. In this phase, the scripts used to produce the candidate outputs are put into regular production where they can augment existing business processes and workflows. The data scientist also takes time to write a formal report that archives and shares the experiment.

As a potential candidate for integration with the intelligence and targeting models, the Data Science Workflow presents several strengths and weaknesses. Two strengths of the model are its iterative nature and its enumeration of specific data science tasks that are performed at each stage of the model (e.g. writing scripts, producing charts, inspecting outputs). The main weakness in the model is that it assumes the model's research questions and other information requirements are previously defined. The model does not include process steps for eliciting or defining what data should be collected or what questions deployed analytics should answer. A common retort to this criticism is that the system will just collect everything and be agile enough to answer any question within the organization's purview. While this may be true, it does nothing to inform the data science team's direction.

ANALYSIS

Our efforts to formally combine the data science, intelligence, and operation processes begin with a functional analysis of the intelligence and targeting processes to identify opportunities to apply data science to cyber operations. The goal of this analysis is to discover the overarching functions that occur in the operations and intelligence processes that drive military cyber operations. Once these functions are identified, we can begin to address how they might be supported by data science. The first stage of this analysis, depicted in Figure 5, is a functional grouping of the common functions found in the intelligence and targeting processes. We identified the functions in each process (the individual cells in Figure 5) and used an affinity diagramming process (Parnell, Driscoll, & Henderson, 2008) to group like-sounding functions into clusters. We then derived a cluster title (the table header in Figure 5) based on the predominate activity that occurs in each cluster (columns in Figure 5). The resultant clusters represent a core set of functions that occur in intelligence and targeting operations. These functions are summarized below.



	Establish Data Requirements	Collect Data	Process Data	Exploit Data	Analyze Data	Disseminate Results	Facilitate Consumption	Gather feedback
Cyber Intel Lifecycle	Requirements	Collection	Process & Exploitation		Analysis & Production	Dissimination	Consumption	Feedback
JP 2-0	Planning & Direction	Collection	Process & Exploitation		Analysis & Production	Dissimination		Eval & Feedback
D3A	Decide	Detect				Deliver		Assess
F3EAD	Find	Fix	Finish	Exploit	Analyze	Dissiminate		

Legend
JP2.0: Joint Intelligence Process
D3A: Decide, Detect, Deliver, Assess
F3EAD: Find, Fix, Finish, Exploit, Analyze, Dissiminate

Figure 5: Functional Analysis of Intelligence and Targeting Processes

Establish Data Requirements. The Establish Data Requirements function is concerned with explicitly specifying what data is needed to inform the rest of the intelligence and targeting process. We believe the notion of requirements, which is enumerated in the Cyber Intel Lifecycle, is a critical function that encapsulates the planning, directing, and decision activities defined in the other intelligence and targeting processes. Requirements represent the outcomes we hope to achieve in cyber operations, and are driven by our national security strategy, our related campaign plans, and the vision of our leaders. These requirements will vary at different echelons. Moreover, they are not fixed, need modification to adapt to changes in our operational and network security environment, and the actions of our adversaries.

Example data requirements include:

- Identify compromised systems within a certain network enclave
- Detect abnormal behavioral patterns by external entities communicating with DODIN assets

Collect Data. The Collect Data function is the act of sensing, storing, and transporting data collected on our networks. This can range from a brute-force approach where everything is collected to a more targeted set of data. The data science team should collect as much data as possible that addresses the requirements without compromising their ability to complete the other steps in the process in a reasonable amount of time.

Process Data. The Process Data function is focused on normalizing, cleaning, and pre-processing the data for follow-on exploitation and analysis. Automated techniques to help translate, classify, and tag the data with machine learning algorithms can markedly aid in this step.

Exploit Data. The Exploit Data function is the act of reviewing the data for analysis opportunities. Analysis can be an expensive and time-consuming process. Therefore, an initial review of the data needs to occur to prioritize where we conduct deeper analysis.

Analyze Data. The Analyze Data function is the application of qualitative and quantitative methods to transform the data into a meaningful result. A meaningful result is defined as one that supports one or more requirements defined in the Establish Data Requirements function.

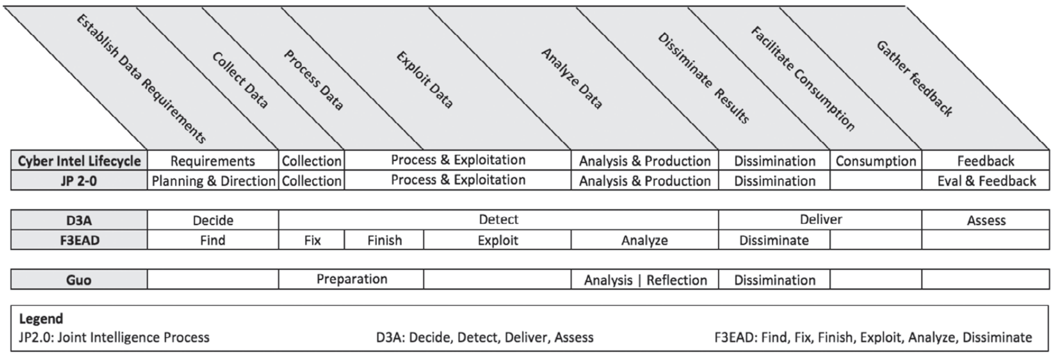
Disseminate Results. The Disseminate Results function is concerned with communicating the results of data analysis with the decision maker, staff officers, other analysts, and curators of the data requirements. This process is quick and continuous in nature. An “always-on” approach aided by artificial intelligent agents that promote and vocalize results can greatly aid in this step.

Facilitate Consumption. The Facilitate Consumption function is a deliberate effort to ensure the disseminated results are consumed to help address data requirements. Of note, only the Cyber Intel Lifecycle included this function. One could argue that this function is not explicitly enumerated in other processes because it is a subtask of dissemination. However, upon further reflection, we believe treating consumption as a distinct function from dissemination is warranted. A data-driven intelligence and targeting process conducted around cyberspace will involve zettabytes of data. As such, there is the potential for a deluge of analytical products, reports, charts, and dashboards that could be produced with this data. Therefore, there needs to be a deliberate and dedicated function to ensure the right analytics get consumed by the right people to make the best decisions. The data science team must devote time coordinating with other analysts, staff officers, and decision makers to understand their workflows and from where they derive their information. The data science team should then tailor and format analytical results to integrate directly with these workflows. This coordination should be done face-to-face.

Gather Feedback. The Gather Feedback function is concerned with working with the decision-maker and other stakeholders to ensure the consumed results of our analysis are actually satisfying our data requirements. This includes verifying and validating both the results and the process used to generate those results. Just because we have a product that reports a certain result, can we trust it?

We next turn to combining the intelligence and targeting processes with Guo's data science process. One approach we considered was simply applying Guo's entire process as a sub-function of each of our eight functions shown at the top of Figure 5. For example, Guo's entire process could naturally nest within the Analyze Data function. Even the Gather Feedback function could benefit from an embedded data science process to help gather and analyze usage data and user behavior. However, we concluded this is a simplistic treatment of data science that will preclude it from reaching its full potential in cyber operations. Data science is much more than a tool or technique to increase the ease and efficiency of our current process. Rather, it is an entirely new approach to how we synthesize, produce, and consume intelligence and operational information in the era of big data. Therefore, we need a more holistic examination of how data science can be integrated with our intelligence and operations processes.

Therefore, we juxtaposed Guo's top level functions–preparation, analysis, reflection, and dissemination–alongside our 8 functional clusters. The results are shown in Figure 6.



	Establish Data Requirements	Collect Data	Process Data	Exploit Data	Analyze Data	Dissiminate Results	Facilitate Consumption	Gather feedback
Cyber Intel Lifecycle	Requirements	Collection	Process & Exploitation		Analysis & Production	Dissimination	Consumption	Feedback
JP 2-0	Planning & Direction	Collection	Process & Exploitation		Analysis & Production	Dissimination		Eval & Feedback
D3A	Decide	Detect				Deliver		Assess
F3EAD	Find	Fix	Finish	Exploit	Analyze	Dissiminate		
Guo		Preparation			Analysis | Reflection	Dissimination		

Legend
JP2.0: Joint Intelligence Process
D3A: Decide, Detect, Deliver, Assess
F3EAD: Find, Fix, Finish, Exploit, Analyze, Dissiminate

Figure 6: Functional Analysis of Intelligence, Targeting, and Data Science (Guo) Processes

This revealed several interesting findings. First, much of Guo’s process lines up well with our military intelligence and targeting processes. However, as previously noted, Guo’s process has no upstream requirements. We also noticed that Guo’s process has no explicit notion of exploitation. It can be argued that exploitation occurs in reflection, but this happens well after analysis so would involve a significant delay. We believe a data science process should feature an opportunity for early exploitation before significant time is invested in analysis. Next, we noticed feedback in Guo’s model is confined to the analysis loop and immediately following the reflection phase. But no feedback occurs outside the process to refine what data gets collected. Finally, Guo’s model does not address consumption.

CYBER DATA SCIENCE PROCESS (CDSP)

> There needs to be a deliberate and dedicated function to ensure the right analytics get consumed by the right people to make the best decisions.

Based on this analysis, we produced a hybrid process we call the Cyber Data Science Process, which is shown in Figure 7. This process model combines the functions from the intelligence and targeting models with Guo’s data science process, building on common functions and addressing gaps. It is extremely important to note that this process is theoretical, and is intended to serve as a conceptual framework for thinking about how to best integrate data science into cyber operations. In practice, the entire CDSP process has to occur within the decision cycle of decision makers; else the entire effort lacks benefit from a military standpoint. Therefore, a data science team may choose to abbreviate, augment, or skip entire portions of the CDSP to accomplish the mission. Our aim is to provide concepts, functions, and terminology to inform the data science team's development of its internal practices.

The CDSP has seven functions, and merges the four major data science functions from Guo, with the functions identified in our functional analysis of the intelligence and targeting process. Each of the CDSP functions are described in detail below.

Establish Requirements. The goal of this function is to establish what data science outputs are needed to ensure friendly force mission accomplishment in the presence of cyber threats and the overall network environment. We emphasize that, at this stage in the process, the focus remains on data science outputs, and not on data science inputs (i.e. what data is required for collection). This is challenging, as it's difficult to envision products and outputs that will result from hours of prototyping, iteration, and testing. However, focusing on what product is needed—a resource decision, identification of a specific target, detection of an enemy operation—will ensure requirements are correctly

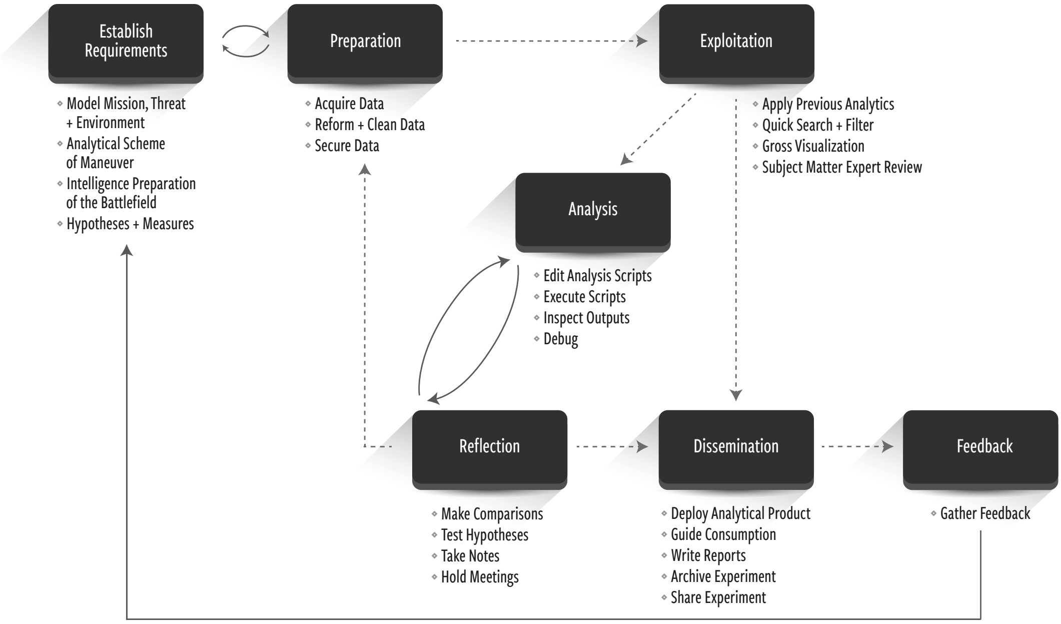


Figure 7. Cyber Data Science Process

established. We propose the data science team first develop models of the friendly force mission, the threat, and the environment. These shape what Parnell and colleagues define as the current state (the "what is") and the desired end state (the "to be") (Parnell et al., 2008). Modeling the mission, threat, and environment produces graphical diagrams, simulations, and mathematical models. This effort might include intelligence preparation of the battlefield process (IPB), an established modeling process integral to lethal military operations (U.S. Army, 1998). Recent work offers a cyber perspective (Winterfeld, Steven P., 2001; Harrison Kieffer, 2016) on the IPB process. Modeling should also include enumerating assumptions, limitations, and constraints relative to the friendly force mission, the enemy, and the environment. The data science team should then specify an "Analytical Scheme of Maneuver" (Stanton, Paul, 2017) to think through the analytical questions, how they relate to one another, how they support the mission, and when and what analytical outputs are needed. The data science team can then form hypotheses that help measure the progress of moving from the current state to the end state. For example, a current state of affairs might involve a suspected threat operating on our networks. The desired end state is the elimination of this threat from the networks. A corresponding hypothesis for this example might entail confirming, or failing to confirm with certainty, the presence of certain network signatures. Once hypotheses are identified, the data science team can draft measures, or metrics, that later confirm or fail to confirm, the hypotheses. These measures directly define what data is required in the next phases of the CDSP. The ultimate product from this phase is a set of formally documented hypotheses and associated data science requirements. Skills required in this phase include data science, cyber operations/intelligence expertise, network systems engineering, mathematical modeling, human behavior modeling, and simulation engineering.

> The Cyber Data Science Process is theoretical in nature, and must be adapted to match the speed and tempo of specific missions and their associated decision cycles.

Preparation. The goal of the preparation function is to identify, collect, and transform all data needed to address the requirements. The hypotheses and associated metrics defined in the requirements function drive what data must be acquired, how much data is needed, and how often it is updated. Preparation includes determining where data collection sensors are placed in the physical word and cyberspace, which entities they monitor, and how data is routed from sensors to persistent storage. Once collected, the data is transformed for follow-on analysis. This is a significant task involving capturing data from sensors and moving and consolidating this data to a persistent data store. A typical cyber security environment can include a wide variety of sensors that collectively produce an extremely high volume of data at high velocities. These range from systems capturing raw network packets traveling at 1000 bits per second to others capturing the hundreds of minute changes that occur on each computer system. These sensors are typically not collocated in the same geographical location, so sensor data is moved to a separate persistent data store where it is available for follow-on analysis. From there, the data is secured to prevent the enemy from manipulating the data to deceive our analytics. The data is then reformatted and cleaned. Reformatting includes actions to store the data in a format that is compatible with the persistent data store. These actions might first include decrypting, decompressing, unpacking, renaming or filtering the original sensor data. Then, data is parsed, translated, and mapped from its native schema into the schema of the persistent data store. The data is then cleaned. Data cleaning, also known as data normalization, is the process of ensuring data integrity and involves deliberate steps to address incomplete, duplicate, or inconsistent records. The ultimate product of this phase is an accessible and persistent data repository containing all the data needed to explore the hypotheses; and, free of error. If this cannot be achieved, the data science team may need to return to the Establish Requirements function. For example, the team may discover they lack sufficient storage or computation resources to prepare the data originally scoped in the Establish Requirements function. Through additional analysis, the team may realize they can accomplish the same requirements with an extract of the original dataset. Skills required for the Preparation function include data science, data architecture, database administration, computer science, information technology administration, and network systems engineering.

Exploitation. The exploitation function involves an initial and rapid review of newly processed information to identify high-value and time-sensitive information that can immediately support the mission. The products of this stage are results from currently deployed analytics, charts, and scripts that immediately answer current or past hypotheses. New scripts or analytical products are not required, and the data science team must resist the urge to launch a new analysis expedition. Instead, the data science team refreshes previously constructed queries and analytics to identify any changes to the status quo or spot obvious items of interest. For example, a simple query searching through network traffic for a discrete set of target IP addresses might return a hit on newly ingested data. The data science team should fully leverage automated systems in this stage to programmatically select, execute, and summarize previously designed scripts and queries. The exploitation phase should also include some level of gross visualization which we define as automated charts and maps that track aggregate trends in the data. Consulting subject matter experts (SMEs) from the cyber mission and intelligence domains is critical in this stage. Their experience and intuition can identify trends and opportunities in the data and refine requirements for follow-on analysis. If the Exploitation function answers the mission-focused data requirements, the data science team can proceed directly to the Disseminate function to share these results. Skills required for the Exploitation function include data science, information visualization, and cyber operations/intelligence expertise.

Analysis. The analysis function of the CDSP involves authoring and editing scripts to test the hypotheses created in the Establish Requirements stage. This function includes an internally iterative process similar to the Analysis phase of Guo's workflow (Guo, 2012). This may involve writing multiple candidate scripts that each attempt to address the hypotheses in different ways. The data science team initially runs and tests these scripts on a subset of data loaded on local computing resources (i.e. a local cluster, server, or workstation). Once tested on an extract of the data, the data science team uploads the scripts into a production data science computing environment such as the Army and Defense Information Systems Agency (DISA) Big Data Platform (Bart, 2016). These environments feature a cluster of scalable computing resources with distributed computing technology such as Apache Hadoop or Spark. These clusters can efficiently apply the newly coded scripts against extremely large amounts of data at Petabyte scale. Even though the new scripts are running on the production environment, they should be designated with a development status until approved for dissemination. Once the scripts are complete, the data science team can inspect their output. This involves examining raw output files and creating visualizations for output data. From these results, the data science team must verify the scripts' output matches intended behavior. If not, the data science team must redesign and debug the script. The ultimate product of this phase is a set of verified analytics (script outputs) that potentially answer hypotheses from the Establish Requirements phase. Skills required in this phase include data science, computer science, mathematics, statistics, machine learning, and information visualization.

Reflection. Once the data science team has a set of validated analytics, they enter the Reflection phase. The goal of this phase is to determine if the hypotheses from the Establish Requirements phases are answered by the analytics. The team makes comparisons and selects which analytics answer the hypotheses in the shortest amount of time with the least probability of error. Documentation and collaboration is essential in this phase, and the data science team should engage the decision maker, SMEs, and other stakeholders to solicit feedback from the newly scripted analytics. If the analytics do not meet the requirements, then the data science team may need to return to the analysis phase and redesign scripts. Or, the team may determine that more data, or data from additional sources is required to answer the hypotheses and return to the Preparation Phase. The ultimate product of this phase is a set of analytics that allows a decision maker to answer the hypotheses. Skills required in this phase include data science, computer science, mathematics, statistics, machine learning, information visualization, and cyber operations and intelligence expertise.

> We believe the CDSP process, which integrates core functions from intelligence, targeting, and data science contains the necessary steps in sufficient detail to guide data science teams as they are integrated into Army cyber operations.

Dissemination. The end products for this phase are permanently deployed analytics running in the production data science environment that are regularly consumed as part of broader cyber operations workflows. In the short term, this involves making the new analytics, previously tagged as developmental, fully accessible to all relevant stakeholders on the production system. The analytics are integrated into dashboards and similar tools and fully documented. The analytics are carefully secured to ensure the enemy does not compromise our data-driven decision-making processes. Additionally, the data science team works to educate and train cyber operators and other stakeholders to adopt and consume the new analytics as part of their regular workflows. In the long term, the data science team should report their efforts to the broader community and archive any results. Skills required during this phases include data science, data architecture, information technology administration, and training.

Feedback. The final phase of the CDSP is feedback. The outcome of this process is a regular review of the deployed analytics' performance, validity, relevancy, and data sources. Performance data—latency, accuracy and resource consumption—is compiled and reported for each analytic. Likewise, each analytics' data sources are reviewed to ensure their integrity. For example, collection may suddenly be interrupted for a data source supplying an analytic which could drastically alter its output. Any issues can prompt a redesign of an analytic. The data science team should also collect and review usage data about how users consume the analytic. A change in consumption could equate to a training deficiency, a loss of confidence in an analytic, or changing information requirements. The original hypotheses are reviewed to ensure they are still relevant to the organization's mission and operations. If these have changed, the entire process is restarted to address evolving requirements. Skills required during this phases include data science, data architecture, and information technology administration. The team should fully leverage automation in this phase to make the consolidation and reporting as easy as possible.

CONCLUSION

In this paper, we outlined the Cyber Data Science Process (CDSP) as a means to guide the application of data science to cyber operations. We believe this process, which integrates core functions from intelligence, targeting, and data science contains the necessary steps in sufficient detail to guide data science teams as they are integrated into Army cyber operations. It is important to remember that this process is theoretical in nature, and must be adapted to match the speed and tempo of specific missions and their associated decision cycles. We also feel this process is implementation and technology agnostic and can be successfully applied at various echelons and across teams of varying size and composition. Moreover, by answering the "how will data science be applied" question, the CDSP helps frame the next set of important questions—who will perform data science in the Army and what capabilities will they need? Toward this end, the CDSP helps specify what skillsets are needed, at what levels, for each of its functions. The process also helps scope task organization options and defines how many soldiers and civilians are needed to keep it running at a particular echelon. It also provides insights into the types of tools and technologies needed in each of its steps. Successfully addressing all of the questions will ensure the Army is well-positioned to realize the promises of data science, increase cyber situational awareness while maintaining information dominance over our adversaries.

The views and opinions expressed in this paper and/or its images are those of the author(s) alone and do not necessarily reflect the official policy or position of the U.S. Department of Defense (DOD), U.S. CYBERCOM, or any agency of the U.S. Government. Any appearance of DoD visual information for reference to its entities herein does not imply or constitute DOD endorsement of this authored work, means of delivery, publication, transmission or broadcast.

NOTES

Baker, J. W., & Henderson, S. J. (2016). Making the Case for Army Data Scientists. *Army,* 66(8), 41–43.

Bart, Daniel V. (2016). Big Data Platform (BDP) and Cyber Situational Awareness Analytic Capabilities (CSAAC). Presented at the AFCEA Defensive Cyber Operations Symposium. Retrieved from http://www.disa.mil/~/media/Files/DISA/News/Conference/2016/AFCEA-Symposium/4-Bart_Big-Data_Platform_Cyber.pdf.

Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., & Simoudis, E. (1996). Mining business databases. *Communications of the ACM,* 39(11), 42–48.

Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (2007). The knowledge discovery process. In *Data Mining* (9–24). Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/978-0-387-36795-8_2.pdf.

Defense One. (2017, January 9). The Pentagon Needs Its Own Google For All Its Data, Says Eric Schmidt - Defense One from http://www.defenseone.com/technology/2017/01/pentagon-needs-its-own-google-all-its-data-says-eric-schmidt/134456/(accessed January 26, 2017).

Faint, C., & Harris, M. (2012). F3EAD: Ops/Intel Fusion "Feeds" The SOF Targeting Process. *Small Wars Journal,* 31(7), 54pm.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine,* 17(3), 37.

Guo, P. J. (2012). *Software tools to facilitate research programming.* Stanford University. Retrieved from http://pgbovine.net/publications/Philip-Guo_PhD-dissertation_software-tools-for-research-programming.pdf.

Harrison Kieffer. (2016). Can Intelligence Preparation of the Battlefield/Battlespace Be Used to Attribute a Cyber-Attack to an Actor? *Cyber Defense Review,* (Spring 2016).

Heidorn, B. (2016, November). *Data Science Panel Discussion.* Sierra Vista, AZ.

INSA. (2015). *Tactical Cyber Intelligence.* Intelligence and National Security Alliance. Retrieved from https://issuu.com/insalliance/docs/insa_tacticalcyber/1.

JP 2-0. (2013). *Joint Intelligence.* Department of Defense.

Juvonen, A., & Sipola, T. (2012). Adaptive framework for network traffic classification using dimensionality reduction and clustering. In 2012 *IV International Congress on Ultra Modern Telecommunications and Control Systems* (274–279). https://doi.org/10.1109/ICUMT.2012.6459678.

Klösgen, W. (1996). Knowledge discovery in databases and data mining. In *International Symposium on Methodologies for Intelligent Systems* (623–632). Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-61286-6_186.

Klösgen, W., & Zytkow, J. M. (2002). Knowledge discovery in databases: the purpose, necessity, and challenges. In *Handbook of data mining and knowledge discovery* (1–9). Oxford University Press, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=778216.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review,* 21(01), 1–24.

Marr, Bernard. (2016, September 9). Big Data In Banking: How Citibank Delivers Real Business Benefits With Its Data-First Approach. *Forbes.* Retrieved from http://www.forbes.com/sites/bernardmarr/2016/09/09/big-data-in-banking-how-citibank-delivers-real-business-benefits-with-their-data-first-approach/#4bed2d1775ed.

Parnell, G. S., Driscoll, P. J., & Henderson, D. L. (2008). *Decision making in systems engineering and management.* Wiley-Interscience.

Reinartz, T. (2002). Handbook of Data Mining and Knowledge Discovery. In W. Klösgen & J. M. Zytkow (Eds.) (185–192). New York, NY, USA: Oxford University Press, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=778212.778241.

Russom, P. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter,* 1–35.

Security for Business Innovation Council. (2012). *Getting Ahead of Advanced Threats.* RSA. Retrieved from https://www.rsa.com/en-us/resources/achieving-intelligence-driven-information-security-synopsis.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing,* 5(4), 13–22.

Sipola, T. (2015). Knowledge Discovery from Network Logs. In *Cyber Security: Analytics, Technology and Automation* (195–203). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-18302-2_12.

NOTES

Stanton, Paul. (2017, March 2). Email Correspondence regarding cyber data analytics.

The White House. (2014). Big Data: Seizing opportunities, preserving values. *Washington, DC: Executive Office of the President.*

U.S. Army. (2010). *The Targeting Process* (No. FM 3-60.).

U.S. Army, U. S. (1998). *Intelligence Preparation of the Battlefield (FM 34-130).* Department of the Army (US).

van der Heijden, T. H. C. (2012). Process mining project methodology: Developing a general approach to apply process mining in practice. *Master of Science in Operations Management and Logistics. Netherlands: TUE. School of Industrial Engineering.* Retrieved from http://alexandria.tue.nl/extra2/afstversl/tm/Van_der_Heijden_2012.pdf.

Verizon RISK Team. (2015). 2015 Data Breach Investigations Report. Retrieved from http://www.isaca.org/chapters2/Luxembourg/Documents/201510%20Digital%20Forensics%20Master%20Class/5_DBIR%202015%20-%20CLUSIL-ISACA.pdf.

Winterfeld, Steven P. (2001). *Cyber IPB.* SANS. Retrieved from https://cyber-defense.sans.org/resources/papers/gsec/cyber-ipb-103147.