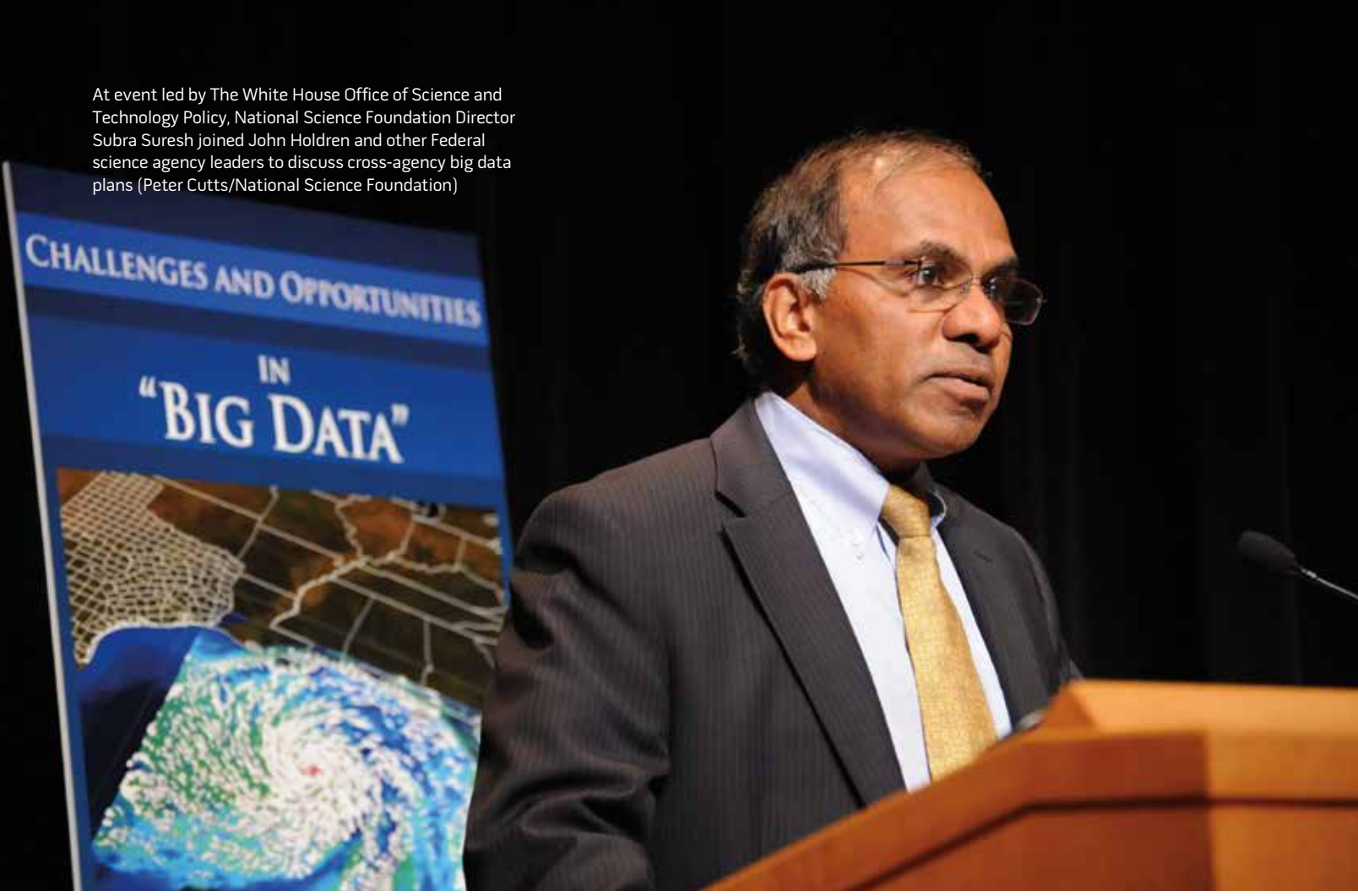


At event led by The White House Office of Science and Technology Policy, National Science Foundation Director Subra Suresh joined John Holdren and other Federal science agency leaders to discuss cross-agency big data plans (Peter Cutts/National Science Foundation)

Framing the Big Data Ethics Debate for the Military

By Karl F. Schneider, David S. Lyle, and Francis X. Murphy

Big data is everywhere these days. It shows up in many realms of contemporary life, ranging from how people are guided to potential purchases as they shop online, to how political campaigns win elections, and even to when farmers plant crops and apply fertilizer to their fields.

While there is no denying the value that comes from data integration and information availability made possible by modern computing power, there are many associated challenges that relate to the privacy of the individual, security of personal data, and reach of decisions influenced by big data. These concerns describe an emerging discipline known as the ethics of big data. This growing conversation is relevant for the military, given both the potential gains from big data collection and analysis as well as the simple fact that big data is here to stay.

In this article, we first define what is actually meant by the terms *big data* and *ethics of big data*, explore the challenges associated with big data, discuss some examples and implications for the military, and conclude with a framework for addressing many of these challenges.

Karl F. Schneider, SES, JD, LL.M., is Principal Deputy Assistant Secretary of the Army for Manpower and Reserve Affairs. Lieutenant Colonel David S. Lyle, USA, Ph.D., is Director of the Office of Economic and Manpower Analysis in the Department of Social Sciences at the United States Military Academy at West Point. Major Francis X. Murphy, USA, is a Research Analyst in the Office of Economic and Manpower Analysis.

The term *big data* or *mega-data* refers to a collection of data sets so large and complex that the data become difficult to process using on-hand database management tools or traditional data processing applications. Big data arises from follow-on analysis of existing large data sets and the capture of software logs and information-sensing mobile devices such as cameras and global positioning systems. E-commerce retailers such as Amazon can exploit data on past Internet browsing histories to deliver targeted, personalized advertising to specific customers. As new technologies emerge and become more affordable to collect, process, and store data, the volume of data collection grows exponentially; some 2.5 x 1018 exabytes of new data are created every day.[1] In addition to the sheer volume of information in big data, these data collection efforts increase the breadth of information available to analysts while compressing the delay between data collection and its subsequent analysis. For example, researchers at the University of Michigan now construct social media indexes of labor market activity such as job loss and job posting using text searches of Twitter posts, a tool that is much more accurate in predicting hiring trends than the consensus forecasts of experts and that is close to being available in real time.

A fairly new and emerging field, the ethics of big data has started to address some of the challenges associated with big data, many of which are of an "ought to" rather than an "is" nature. Big data itself is ethically neutral; it is the actual *use* of big data that raises ethical questions. Thus, the ethics of big data concerns more than simply the matter of morality. Rather, it includes issues such as the privacy, validity, security, transfer, and analysis of big data as well as the business decisions or policy implementation that follow from big data insights. These topics have far-reaching consequences when the data relate to sensitive homeland security matters, individual medical records, or more broadly to data containing personally identifiable information, which often include sensitive information such as name, date of birth, and Social Security Number (SSN).

Understanding Big Data and the Ethics of Big Data

An important starting point for understanding big data is to consider the structure of the underlying information. Big data is referred to as structured when it is in traditional rows and columns such as one would find in a standard spreadsheet. At the other end of the continuum, photographs or feeds are considered unstructured data. Free-form text in a social media status update is an example of semistructured data and sits at the middle of this continuum because it has features similar to both structured and unstructured data.

One of the most appealing aspects of big data and its applications is the ability to study a larger share or sample of an underlying population. Large samples allow researchers, policymakers, and business analysts to better approximate how behavioral responses vary across different segments of an entire population. They also allow for increased understanding of heterogeneity, or granular differences across variables within data. For example, in figure 1, we compare two samples of different size drawn from the same underlying population. Having access to a larger sample with more granular, accurate, and timely data—such as in Sample B—allows for a more complete analysis of behavior among those in the underlying population.

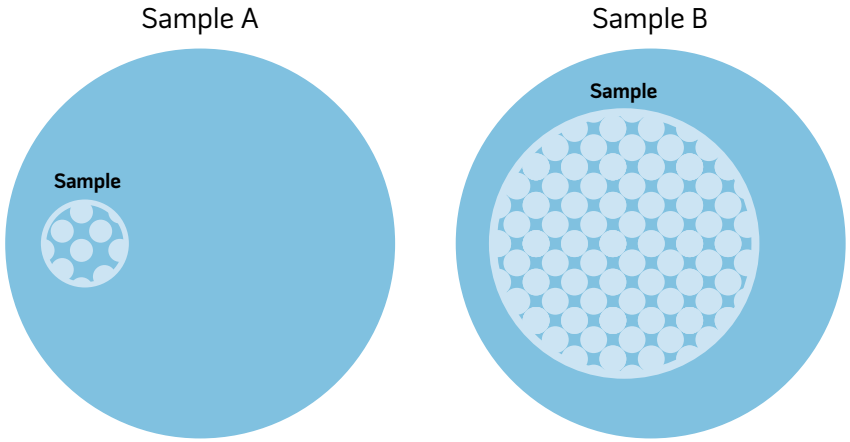
There is a growing dialogue aimed at formalizing norms related to the ethics of big data. The White House and Massachusetts Institute of Technology recently cohosted a forum titled "Big Data Privacy: Advancing the State of the Art in Technology and Privacy." In April 2014, the University of Virginia hosted the first "National Conference on Big Data Ethics, Law and Policy." The Council for Big Data, Ethics, and Society also convened in 2014 to address security, privacy, equality, and access to big data. Standards related to the collection and use of big data are the focus of an emerging field of study; in *Ethics of Big Data: Balancing Risk and Innovation*, Kord Davis advocates for a framework based on identity, privacy, ownership, and reputation.[2] He believes that asking questions along these dimensions will help to establish common ground for discussing the ethics of big data. Similar to the work by Davis, a law review article by Neil Richards and Jonathan King suggests that principles of privacy, confidentiality, transparency, and maintaining identity should govern data flows and inform the establishment of big data norms.[3] Finally, the White House recently released a major study on big data that emphasizes, among other things, the need to preserve personal privacy even as the promise of big data suggests better delivery of nearly every type of public good.[4]

These early efforts to address the ethics of big data have helped elevate the importance of this topic and have highlighted the need for a common language and clear guidelines that promote understanding of expectations and best ethical practices. Implicit within the area of big data ethics is an application of common practices to each specific institution. For example, the issue of trust may have greater weight for an organization such as the military than for an organization such as Facebook. This is not to suggest that trust is not important in the private sector (because it certainly is); however, trust does have an amplified importance in the military because life and death outcomes as well as national security are at stake.

Benefits of Big Data

As discussed, big data has many practical uses in fields such as research, disease prevention, and business analytics. Examples of big data include Google Analytics on the frequency of Internet searches to predict share price movements of publicly traded companies and the Sloan Digital Sky Survey, which collects 200 gigabytes of astronomy-related data per night. eBay and Facebook also maintain almost unfathomable quantities of data on consumer transactions and user-uploaded photos, respectively. Big data collection of consumers' Internet browsing and purchasing histories helps e-commerce firms such as Amazon build algorithms for applications such as its "You Might Like" feature. The retailer here relies on big data to recommend products

Figure 1. Big Data's Impact on Sample Sizes



that the Internet shopper is likely to purchase.

Big data is also relevant to decision-makers in fields as diverse as agriculture and political campaigns. Big data analysis of variation in soil fertility and nutrient needs within a field—coupled with global positioning technology—now allows farmers to customize the application of fertilizers, reducing input costs while increasing yields. Since 2004, political campaigns have used big data for "micro-targeting"—personalizing campaign-messaging based on voter demographics. Campaign strategists increasingly use big data to identify the most efficient uses of campaign funds to persuade prospective voters to go to the polls and vote for their candidate.

An example of big data's potential benefit to the military is the ongoing empirical research measuring the impacts of military service on lifetime earnings. This project requires the development of a unique data set containing individual-level structured data on millions of veterans across nearly two decades. Through a series of data merges, administrative data are joined with information on veterans' disability compensation, GI Bill usage, as well as unemployment benefits receipt and annual earnings.

Big data allows the researcher to combine and subsequently analyze all of the above dimensions of a veteran's experience. Thorough analysis of this data by skilled researchers can identify the long-term impacts of military service and the use of specific military benefits. While the data effort is immense, the proposed research outcome is critical to society and its understanding of the costs and benefits of the all-volunteer force. This is just one research-driven example of the power and utility of big data when it is harnessed properly.

Although the above example illustrates the potential value of big data to the military, it also highlights the larger ethical implications of assembling big data. First, the collection and use of big data cannot compromise the organization's core value of trust: that the military will both provide for the national defense and also look out for the best interest of its Servicemembers. The military must recognize that the individual has enduring rights related to personal information, regardless of whether a third party or agency has access to or custody of that data.[5] Second, as we address in more detail later, it is crucial to realize that big data itself is not a cure-all as some have suggested. Rather, big data is complementary to existing methods as a high-powered analytical tool that nonetheless requires the context of understanding the problem, considering theory, formulating hypotheses, and testing for relationships.

Challenges of Big Data

While the opportunities presented by big data are immense, the associated challenges are important and must be considered by any person or organization involved in the collection and analysis of big data. Accordingly, we group challenges associated with big data into six principal areas:

- use and transfer of personally identifiable information (PII)
- merging and combining data sets
- policy formulation at the individual and group levels
- costs associated with use and analysis
- personnel challenges
- general analysis and interpretation.

The first challenge comes from the use and transfer of PII, which is information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context. The big three identifying variables are name, date of birth, and SSN. Together, these three pieces of information can almost always allow individuals to be identified within the data. There are some important distinctions concerning the nature of this information. PII can be sensitive or insensitive in terms of the consequences of its release or use. It can also be voluntarily or involuntarily collected from the subject, and the subject may be aware or unaware that the data have been collected or used. Finally, the data may be required by the military, as in the case of information on health conditions, or the data may be extraneous resulting from recent browsing history, for example. These distinctions will dictate, for instance, the level of security required for storing or transferring the data and whether the individual should be informed about potential uses of the data. It is also important to realize that these distinctions are dynamic and that the custodian of the data has to be constantly aware of the changing nature of data.

One of the key attributes of big data is that much of it follows from the integration or derivative analysis of data that already exists. The project on veterans' lifetime earnings is an example. Whenever big data is generated through the combination of existing data sets, there is the potential that previously anonymous data can become identifiable for an individual as a result of being merged with other

data. The new and larger data sets that result from combining data like this almost certainly rely on unique identifiers such as SSNs for accurate merging, and what results may have a new level of sensitivity. For instance, combining For Official Use Only (FOUO) information on the members of a military unit with FOUO information on unit locations in contingency operations will lead to a larger data set that is now classified as Sensitive. A related concern is the privacy right of the individual when a custodial entity is merging and sharing data. The purpose of the merge and the nature of the data dictate whether the subjects need to be aware of the merge or perhaps must even give permission before merging. Moreover, the riskiness of the data-sharing increases as it gets farther away from the source; merging Department of Defense (DOD) data with other DOD entities presents less risk than merging that same data with private corporations. Whether merging inside or outside of the military, deliberate care must be taken to mitigate the amount of information that merging entities gain. Many straightforward encoding techniques for merged data are available that can significantly reduce what is actually shared with the outside party.

Next, there are ethical implications concerning the use of big data analysis for policy at the individual or group level. Policymakers must be careful about how they use insights derived from big data, whether that data are unwillingly or willingly provided by the subject. Servicemembers have a right to privacy, and it can be problematic for an organization such as the military to use an individual's data against him.[6] For instance, imagine that the military conducts an analysis of Sergeant Smith's use of medical care benefits. If DOD enacted Service-wide policies based on that analysis, the ethical concerns are minimal. Suppose, instead, that one of the Services targeted Sergeant Smith with new premiums based on the data analysis. Given the unique mission and culture of the military, this example of micro-targeting would likely be viewed as a breach of trust and hence an inappropriate use of big data. In fact, such initiatives could engender resentment, lead to unintended changes in healthcare use, and even provide individuals with a strategic incentive to misreport their preferences for healthcare services as well as actual use.

Thus far, we have focused on the security challenges inherent in assembling big data within the military. Effective use of big data also requires *time*, *talent*, and *money*, all of which are scarce resources within any organization.

Identifying the necessary data and then constructing big data take significant effort and time. For example, the U.S. Army's personnel database has more than 2,000 variables per observation. However, a personnel analyst might routinely use only 200 or fewer of these. Extracting the relevant variables from this larger data set and then preparing them for analysis are time consuming and require individuals with both institutional knowledge of the Army and expertise in database management. Coordinating data merges with other government agencies is a lengthy process as well, and again requires individuals with expert knowledge. Moreover, collecting, storing, and safeguarding big data can require costly investment in state-of-the-art infrastructure and security software. The costs associated with training personnel to use big data, and subsequently providing sufficient tenure to these personnel so that the military can recoup the investments in developing this institution-specific knowledge, are substantial. In fact, this tenure problem may be one of the biggest challenges for the military, since the existing promotion system encourages frequent job-switching.

Big data reinforces the need for internal control mechanisms, such as institutional review boards (IRBs), which provide important oversight. However, the military has historically confused Privacy Act requirements and Human Subjects requirements with its IRBs. Human Subjects protections apply to some, but not all, information protected by the Privacy Act. Human Subjects protections are enforced by IRBs with the purpose of ensuring that a study is conducted ethically and without imposing any harm on an individual. These protections are designed to prevent abuses of human subjects in *experiments* and do not apply to policy analysis of existing administrative data. This is a key distinction that leadership must appreciate and support.

Similarly, big data implementation raises personnel challenges. The military must be deliberate in selecting and training personnel who use big data. Big data requires server and storage hardware, software, system administrators, database administrators, and analysts; each of these jobs requires specific skill sets. The development and continuous maintenance of those skills entail significant investments. DOD must also determine levels of access for anyone who interfaces with big data. Moreover, each class of military personnel (enlisted, officer, civilian, contractor) that works with big data not only provides unique benefits but also presents unique challenges and risks. Reiterating, extended tenure is a necessary condition for analysts using big data.

Finally, many challenges are related to the general analysis and interpretation of big data. There is an overriding temptation to equate the sheer quantity of the data with the mistaken assumption that any findings from such massive data must be meaningful. This belief is particularly dangerous since big data—based on sheer sample size—tends to produce many statistically significant findings, even if the proposed relationships are spurious and the analytical methods inappropriate. Regardless of the volume of data, analysis must be guided by relevant theory and sound statistical inference. In other words, big data must be paired with *big judgment* for the analysis to have practical policy applications or business decision relevance.[7] The often repeated mantra in the social sciences, "correlation does not imply causation," certainly rings true here. The popular economics writer Charlie Wheelan addresses this distinction in his 2013 book *Naked Statistics*, in which he imagines a study linking 5- to 10-minute outdoor breaks taken by office workers to increased rates of lung cancer. Of course, it is not the outdoor breaks that are causing cancer, but rather the smoking of cigarettes while outside on

Results of bacterial susceptibility tests were fed into computer and used to create printouts of data showing worldwide patterns of bacterial resistance to antibiotics (FDA)

the break that is the causal factor.[8] Thus, analysts must understand underlying sources of variation and consider inter-variable relationships so that they can differentiate between true causal relationships and mere correlations.

Illustrating the Challenges of Big Data

The Commander's Risk Reduction Dashboard (CRRD) is a current big data application in the Army that illustrates several of the big data challenges outlined above. The Army is increasingly integrating a variety of personnel data and relying on analysis of it to inform decisions at local command and higher levels. Launched in January 2014, the CRRD consolidates information from multiple sources—including medical records, deployment data, and correction actions—to provide unit commanders current snapshots of personnel who might be at high risk of manifesting suicidal behaviors. The CRRD takes the form of a software application that commanders access through Sharepoint, and it represents one effort the Army is using to address the recent increase in suicides among its ranks.

Much of the data in the application (and others that are similar) are personally identifiable and thus sensitive in nature, so there are immediate concerns about security, proper use, and general privacy and identity of individuals. Specific to the CRRD, there must be clear policies for access and transfer of that data: who needs to see the data outside of the command team and through what media? What happens when a Soldier transfers units? A related concern is whether the Dashboard program should apply algorithms to the data and make predictions, or simply allow the commander to observe and then process the raw data himself. Similarly, there is a chance that this exercise in statistical risk projection (whether done by the algorithm or the commander) could lead to prejudgments about a Soldier's performance and potential, particularly in the case of a Soldier trying to make a fresh start in a new unit. What if that Soldier misses out on a promotion, key assignment, award, or superior evaluation because the algorithm has determined that he is at risk for suicide-related behavior? Is this outcome fair? Does it violate the Soldier's right to privacy? Will uninformed use of this data actually increase the Soldier's risk of self-harm? Does the military's prerogative to prevent suicides—arguably at any cost—override these concerns about privacy and fairness? These and other questions capture many of the current dilemmas associated with the use of big data.

A Framework for Big Data in the Military

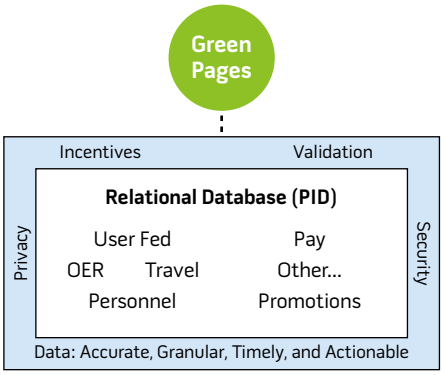
Given the challenges associated with big data and the striking relevance of those challenges to the Army in applications such as the Commander's Dashboard. In this framework, military big data is bound by pillars of privacy and security while incentives and validation result in data that are accurate, granular, timely, and actionable. Firewalls protect data so that authorized users can access, analyze, and use this data in a secure environment. Numerous privacy considerations are a primary feature of the framework and manifest in control measures including but not limited to SSN to employer identification numbers conversions where appropriate, system of records notices, and privacy impact assessments. Complementary to the privacy emphasis are systems ensuring security via controlled data usage and disclosure. These include mechanisms such as IRBs and data use agreements, nondisclosure agreements, and data-sharing agreements. The accuracy, or content quality, of the data comes from having the right incentives and validation processes in place so that analysts and policymakers can be confident in the reliability of what is on hand. For instance, the prerogative to update some dimensions of personal data must be tied to an important outcome for the individual, such as a greater likelihood of promotion, benefit receipt, or future assignment.

Applying the Firewall Framework

The Green Pages pilot program, part of the Army's talent management initiative, is but one example of the potential for the Service to collect, analyze, and use big data to improve officer productivity and satisfaction. The key ideas embedded in the firewall framework were essential to its development and recent Green Pages piloting efforts. Green Pages is a concept that uses a software platform that allows officers to supplement existing administrative records with user-fed data such as hobbies, past experiences, interests, and preferences in the context of seeking out a best-fit assignment. At the heart of the Green Pages concept is the imperative to incentivize and validate the secure collection of accurate, granular, timely, and actionable data; this enables the Army to learn about its on-hand talent inventories while finding optimal employment for its officers. The big data nature of the Green Pages concept cannot be understated: just as the veterans' earnings project discussed earlier combines a variety of existing data, the Green Pages concept integrates administrative data from the Total Army Personnel Database, performance information from officer evaluation reports, and information on civilian schooling and precommissioning experiences. Combining this user-fed information with existing administrative data helps the Army develop a much more complete picture of officer preferences and talents.

For the Green Pages concept to be successful, stakeholders must trust that the organization will keep the data secure and respect individual privacy concerns. These initial imperatives correspond to the key concepts of the firewall framework and thus the entire database sits at the center of the firewall, pillared by privacy and security (see figure 2).

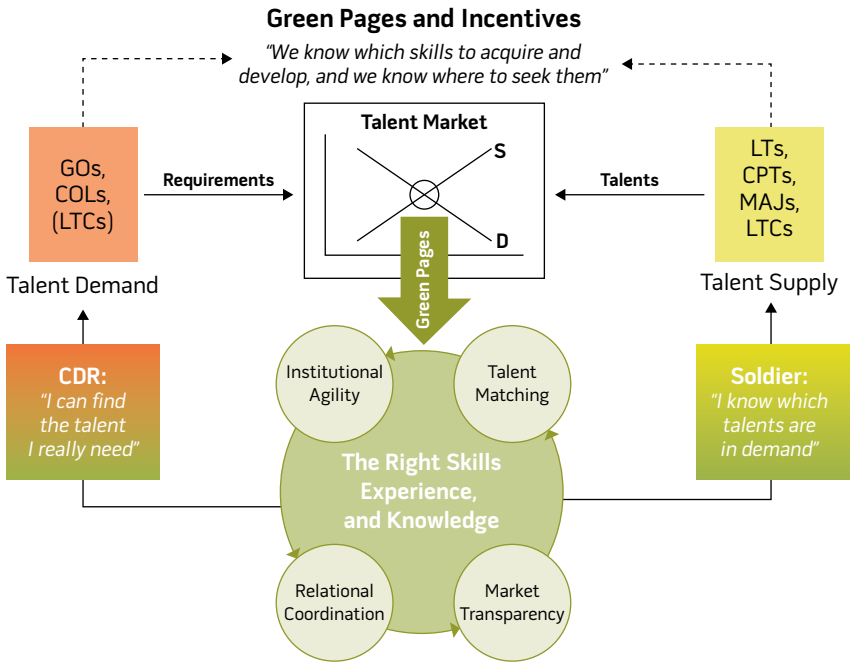
Figure 2. Applying the Firewall Framework to Green Pages



Security is achieved through several features of the Green Pages pilot program. The Web and database servers are configured according to DOD baseline security standards. The servers also use antivirus, host-based firewalls and the DOD Host Based Security System to mitigate known security threats. All Web-based client and server communication is accomplished using appropriate encryption. This secure traffic passes through a reverse Web proxy to conceal the underlying network architecture from potential attackers. A firewall mitigates both injection and scripting attacks. These servers reside within a DOD Data Center with security guards, limited access controlled by Common Access Card (CAC), and security cameras. The database is backed up nightly to network storage. A limited number of system administrators and database administrators have access to the database server and database backups on network storage.

Privacy is achieved through strict authentication procedures and a continuum of what is available at each echelon of profile view. The Green Pages pilot implements authentication using enterprise single sign-on via the Army Knowledge Online (AKO) Single Sign-On. Users gain authentication mostly through CAC/personal identification number with limited use of AKO username/password, where username/password authentication only allows an individual to view her own data. Authorization is controlled by logic built into the Web application. Data access is segmented by role (Headquarters Department of the Army [HQDA], Human Resources Command [HRC], unit strength managers, and users). Within the Web application, users can modify their privacy setting to control the portions of their user profile viewable by others. This does not limit access to their profile by users with higher level roles (HQDA, HRC). At the least restrictive level, strength managers and senior leaders can view basic information such as home address, security clearance, and last change-of-duty station date. Further authorizations enable the user to see the experience overview, current and past chains of command, education, and assignment preferences. Only at the most restrictive echelon can the viewer see full name, email address, current assignment, and organization. Finally, the Green Pages database contains limited PII. To reduce the risk of accidental disclosure, the SSN

Figure 3. Talent Market Enabled by Thoughtful Data Strategy



is replaced with a unique employee identifier, such as AKO username or DOD Electronic Data Interchange number. Similarly, date of birth is replaced with year of birth.

Incentives to maintain the quality of the data are also important to the Green Pages concept; the officer must have "skin in the game" to reveal new elements of the officer's talents and preferences and validate existing information. Whereas in the past officers would receive assignments from the Army without submitting input beyond just their preferences, there is now an incentive structure through the Green Pages market mechanism that is centered on the productive dimensions—or talents—of the officers. We depict this market mechanism in figure 3. Company-grade and field-grade officers represent the supply of talent in this setting, while units represent the demanders of these talents. Optimal talent matches require that both sides of the market have ready access to information that will facilitate matching the right officer to the right position. Within Green Pages, officers (the supply side) are incentivized to reveal their talents while units have a clear incentive to explicitly convey their talent demands. With the proper incentives in place to foster an exchange of accurate, granular, and timely talent data between market participants, the Army collects much better talent data on its officers. Remember that the emphases on security and privacy undergird this entire market information exchange.

Validation is an important feature of the mechanism design for the Green Pages concept. The spectrum of talent-related data is naturally broad because it encompasses anything that measures the officer's potential for productivity. A market-based construct such as Green Pages gives users near instantaneous access to their data. This facilitates validation, corrections, and high-frequency updates to Army administrative data. The fact that peers and mentors can view officer profiles provides an additional important validation mechanism. The linkage between data contained in the Green Pages concept and the assignment process provides perhaps the most powerful validation mechanism of all. For instance, it would not be good for an officer's career if he were hired as a Chinese language expert when the officer's proficiency is actually in Arabic. Thus, the officer has a strong incentive to regularly monitor data in Green Pages and update any information that is incorrect; the result for the Army is validated talent data that is more accurate, granular, timely, and actionable.

A Way Forward

Our nation's military must embrace the fact that big data is here to stay. It must identify methods to tap the vast informational content resident in big data to meet our national security objectives more effectively, while avoiding the negative consequences of uninformed and improper use of this data. Given the importance of trust within these institutions, the stakes are particularly high; misguided uses of this information and potential security breaches of individual data can degrade the morale of Servicemembers and erode public trust in the military.

This article has defined big data and provided a framework for thinking about the ethics of big data in the context of the military. We believe that the firewall framework is one way to orient big data efforts toward security and privacy while incentivizing the provision of accurate and granular data; the firewall provided guiding principles that were useful in the Green Pages case study outlined above. Nonetheless, the firewall framework is simply an initial step in a new area that is still developing and relevant to the military and many other institutions and organizations. Institutions that deal with big data must be mindful to build agility into their formal promises as this is an area of constant change.

Given the vast amounts of data that the military maintains and the high stakes associated with preserving it and ensuring its proper use, the military must engage both internally and externally in the ongoing early dialogues related to the ethics of big data. The military can help to shape the development of the ethics of big data—which will eventually grow into a set of norms with far-reaching implications for both the private and public sectors. Additionally, the military needs to have thoughtful protocols for securing, transferring, storing, and using big data, and must update these protocols with changing technologies. Moreover, the military

GEN Ann Dunwoody, Commanding General of Army Materiel Command, and Mrs. Linda Via promote LTG Dennis Via to rank of general during ceremony at Redstone Arsenal, Alabama, August 2012 (U.S. Army/Teddy Wade)

must continue to refine systems that ensure proper permissions are requested and granted for accessing and using big data. Finally, it must educate the force not only on the proper ethical use of data, but also on the correct use of statistical procedures used to inform decisionmaking. This training must extend to the consumers of analysis so that the military can implement appropriately informed policies. This type of training is crucial for the leaders of tomorrow's military and would be an appropriate feature at all levels of professional military education. **JFQ**

Notes

[1] "What Is Big Data? Bringing Big Data to the Enterprise," *IBM.com*, available at <http://www.ibm.com/big-data/us/en/>.

[2] Kord Davis, *Ethics of Big Data: Balancing Risk and Innovation* (Sebastopol, CA: O'Reilly Media, 2012).

[3] Neil M. Richards and Jonathan H. King, "Big Data Ethics," *Wake Forest Law Review*, January 2014, available at <http://ssrn.com/abstract=2384174>.

[4] President's Council of Advisors on Science and Technology, *Big Data: Seizing Opportunities, Preserving Values* (Washington, DC: The White House, 2014), available at <www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf>.

[5] Ibid., 32–34. See the brief discussion of foundational "third-party doctrine" Supreme Court cases and associated statutes, such as the 1974 Privacy Act.

[6] Ibid., 51–53. The recent White House report cautions against the use of big data analysis to enable discrimination such as in this example.

[7] Shvetank Shah, Andrew Horne, and Jaime Capella, "Good Data Won't Guarantee Good Decisions," *Harvard Business Review*, April 2012.

[8] See Charlie Wheelan, *Naked Statistics* (New York: Norton, 2013).

We thank members from the Office of Economic and Manpower Analysis at the U.S. Military Academy for their valuable comments and suggestions to this article. The Office of Economic and Manpower Analysis has maintained one of the Army's single largest databases for more than 30 years and has extensive experience in working with big data.