HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

POLICY PAPER
belfercenter.org/homelandsecurity

HOMELAND SECURITY PROJECT | MARCH 2021

Improving Big Data Integration and Building a Data Culture for U.S. Border Security

Stephen Coulthart and Ryan Riccucci

Executive Summary

The potential impact of big data for border security is, in a word, transformative. However, U.S. Customs and Border Protection (CBP) does not leverage much of the large volume of data it collects. If CBP could better integrate these data into operations, it would speed up cross-border trade by helping authorities identify the most at-risk travelers and cargo. Big data can also help policymakers better understand the extent to which the border is secure and improve the allocation of enforcement resources.

Significant barriers remain for CBP to leverage big data, such as information sharing barriers between operational components as well as safeguarding data from breaches. These barriers are caused by a variety of factors. Like much of the U.S. government, CBP has struggled to develop a data culture receptive to changes brought on by the information technology revolution and has significant issues with its data governance standards, technology acquisition, and human capital development processes. This article explores these issues and offers recommendations to address these and other barriers to unlock the potential of big data for U.S. border security.

Background

Over the last 20 years, CBP—the primary U.S. government organization charged with securing the nation's borders—has spent billions on technology to monitor international boundaries, vet international travelers, develop trusted traveler programs, and inspect cargo at ports of entry. These systems are integral to the U.S. government's post-9/11 risk assessment and mitigation strategy.[1] Under this approach, data and computer algorithms are used to categorize travelers and cargo based on the level of potential risk they pose. The riskiest entries—for example, a traveler who repeatedly visits a country with high levels of terrorist activity—are subjected to additional scrutiny, such as advanced inspection.

The vast data collection infrastructure built by CBP contains sensors and computer networks, which hold what can easily be described as big data, that is, "data so large, varied and dynamic that conventional hardware and software cannot process it."[2] Collecting vast amounts of data alone provides little benefit to frontline personnel and senior decision makers—the data must still be operationalized. Consider, for example, the U.S. Border Patrol's Automated Biometric Identification System (IDENT), which holds biometric data on illegal crossers, such as their finger prints. The system is useful because it allows agents to identify repeat crossers and those with previous criminal records. IDENT's value became more apparent when combined with a framework to help agents determine the level of sanctions to apply to illegal border crossers—the Consequence Delivery System (CDS). By fusing data from IDENT with CDS decisions, officials could see how less severe sanctions (e.g. voluntary return) and more severe sanctions (e.g. prosecution) impacted re-apprehension rates, and potentially redirect their enforcement resources accordingly.[3] As this case illustrates, collecting large amounts of data is not enough to draw value from it; it has to be processed and analyzed. Researchers have identified several ways organizations exploit big data.[4] These include, among others:

- The connection of multiple large datasets internal and external to an organization
- The combination of large datasets containing structured data (e.g. data contained in rows and columns, such as a spreadsheet) with unstructured data (e.g. video and photos)
- Processing and analysis of incoming streams of data in real or near real-time (e.g. to create data dashboards)
- Using advanced analytics and artificial intelligence techniques like machine learning to analyze data at scale

Big data capabilities can help CBP find new opportunities to further speed up legitimate trade and travel. Radio-frequency identification-enabled ID cards allow crossers in trusted-traveler programs to move through ports of entry with speed and ease. Cameras at border crossings collect information on travelers' vehicles via license plate readers and, increasingly, facial recognition technology. When these data collection sources are analyzed with algorithms that provide risk assessments, officers can focus their time on entries that pose the greatest likelihood of malicious activity. Meanwhile, legal travel and trade can flow without need for intensive searches, thereby increasing economic efficiency. Similar technologies promise to speed up cargo inspections as well, such as the Automated Commercial Environment, which is a system that streamlines cargo processing.[5]

Maturing big data capabilities can improve situational awareness and help answer the difficult question: "How secure is the border?" As one senior official stated, "We call it the denominator problem. How many people are coming across the border [and] how do we really know how [well] we're doing . . . ?"[6] The benefit of answering these questions is that the U.S. government can more effectively allocate resources to the most porous sections of the border and, ultimately, improve U.S. border security. Away from the border, big data can improve the targeting of criminals by integrating a wide variety of datasets, such as telephone number databases, Western Union wire transfers, and social media posts. The use of these various sources marks a transformative moment in law enforcement—previously investigators were reliant on human sources for leads but now can draw from a host of digital data sources for tips. The use of these data to proactively target vulnerable components of criminal networks also marks a shift towards a disruption strategy and away from a more reactive approach.[7]

In preparing this article[*] we identified many "pockets of excellence" where advanced data analytics are now used.[8] The National Targeting Center, an organization in CBP that screens and analyzes cargo and traveler data, has significant capabilities to fuse and analyze information from a variety of sources. Another initiative at CBP's sister agency U.S. Immigration and Customs Enforcement, the Border Enforcement Analytics Program assists Homeland Security Investigations (HSI) "to effectively combine and analyze multiple, large disparate data sets to increase enforcement effectiveness."[9] CBP's Office of Field Operations, which manages the nation's ports of entry, is also experimenting with new data-intensive processes to screen travelers. Still, most of the personnel interviewed from senior to frontline officials noted CBP is struggling to use the large amount of data it already collects. As one Border Patrol supervisor noted, "We don't have a great method to really analyze or even retrieve [data]. . . . We just dump information to the systems, but then the ability to regurgitate that information to where it's something useful doesn't really exist."[10]

Observations like the one above on the difficulty of sharing data make sense in light of the disconnect between data collection, storage, and distribution processes. Former CBP deputy

[*] We interviewed almost two dozen CBP personnel and experts for this paper. Interviews were conducted with the understanding that responses would be anonymized. Additionally, Institutional Research Board (IRB) controls were in place to protect interviewees and maintain anonymity of their responses.

commissioner David Aguilar noted: “We have unattended ground sensors, we have integrated fixed towers. . . . Each one of those platforms . . . operates independent of each other.”[11] The variety of data streams, organizational missions, and bureaucratic complexity has led to a labyrinth of databases: the Department of Homeland Security (DHS), of which CBP is a part, maintains approximately 900 largely unconnected databases.[12] Studies suggest that up to 60 percent of data held by government agencies are unknown and/or unused—the same is probably true of DHS’ and CBP’s systems to a similar extent.

Other issues arise from the techniques used to analyze big data. Data mining algorithms help authorities sift through large datasets but raise a host of ethical and legal questions.[13] These can include biased assessments when the data, algorithms, or both are improper for a task. For example, the use of algorithms to assess the potential of a person to engage in political violence using their social media activity is not an effective use of data mining. A group of 53 machine learning experts commenting on the proposed “Extreme Vetting Initiative” noted that the characteristics of terrorist and criminal behavior “are difficult (if not possible) to define and measure[, and] any algorithm will depend on ‘proxies’ that are more easily observed and may bear little or no relationships to the characteristics of interest.”[14] A 2020 report from the Brennan Center found CBP used similar algorithms to assess risk of travelers on the basis of their social media activity.[15]

As CBP struggles to leverage its data, it has at times struggled to secure it. In 2019, a contractor tasked with managing a CBP pilot project to collect facial recognition data at ports of entry was the target of a ransomware attack. The images of approximately 184,000 travelers were stolen in the subsequent breach and at least 19 were leaked to the dark web. A DHS Office of the Inspector General report concluded CBP’s information security protocols were insufficient. In particular, a subcontractor downloaded data improperly, which made the breach possible.[16] Another found security vulnerabilities on CBP systems and workstations.[17] As CBP collects more data in the coming years, safeguarding it will be of paramount importance.

Analysis

The problems detailed above arise from a variety of factors:

1. **Organizational barriers limit efforts to share data and best practices across CBP.** The sheer size of CBP—it employs more than 60,000 personnel—along with the diversity of its missions from inspecting agricultural shipments to patrolling the border, leads to barriers to collaboration. A CBP official we interviewed noted that the organization "needs to still grow in terms of how . . . operations interact with each other." [18] Studies on innovation in government suggest breaking down these barriers is a prerequisite for the effective use of big data.[19] In particular, these barriers make information sharing difficult and hinder the ability for different parts of the organization to learn from one another. These challenges often are more difficult when considering information-sharing opportunities between CBP and other agencies, especially where agency objectives are not in alignment and there are differing privacy and civil rights and liberties standards.[20]

2. **Human capital issues hinder the development of new big data capabilities.** While the lack of data science experts—the specialists with the skills and knowledge to leverage big data—is a problem in the wider economy, the lack of expertise is more notable in government. This is true at CBP, although not evenly across all organizations. In some parts of CBP, such as the National Targeting Center, there appears to be a significant amount of data-savvy personnel. In other offices of CBP, such as the Border Patrol, there is far less expertise. One Border Patrol Agent summarized the problem well: "we don't even need data scientists . . . we haven't even fully used Excel yet."[21] The development of data science human capital is important because these specialists play a key role in designing, implementing, and using big data applications. For example, an Office of the Inspector General report of recent technology upgrades concluded the Border Patrol "had inadequate personnel to fully leverage surveillance technology or maintain current information technology systems and infrastructure on site."[22] Adding to the problem, the U.S. government struggles to retain talented employees who can find more lucrative employment opportunities in the private sector. Basic data literacy is also important to ensure personnel understand how to process and interpret data they come in contact with.

3. **CBP struggles to inculcate the values of a data-centric culture across echelons of the organizations.** Implementing big data requires addressing obstacles related to organizational culture.[23] There are several different organizational cultures in CBP. However, more if not all of the various sub-cultures of CBP are not "data cultures," meaning those that encompass "values, behaviors, and attitudes . . . that promote and enable use of relevant data

as the driving force of decision making."[24] A Border Patrol agent interviewed for this article summarized well the difficult transition to a data-driven culture in his part of the organization: "It's going to be a huge paradigm shift from the way we operate now. A perception of work tools and agency and the people we have [are] knuckle draggers [and] ground pounders . . . guys who get in the dirt and go get things done."[25] Making the transition to a data culture is important for a couple reasons. First, an organizational culture that emphasizes data increases intra-organizational information sharing; employees feel safe to share data with co-workers, which benefits all. Second, when data is incorporated throughout an organization, it stimulates bottom-up innovation among employees, an important source of innovation and new prototype technologies to exploit big data.

4. **The government's acquisition process stifles big data innovation.** Acquiring new technologies, ranging from camera systems to data mining software, is necessary for collecting and making sense of large datasets. Back-end technologies, such as server architecture, that transmit data to and from operators in the field are critical to leveraging big data. However, the current approach to acquiring these new technologies in CBP and DHS is complex and slow.[26] Once a technology need is identified, the government writes specific requirements with little input from agents and officers in the field. Since the requirements are highly specific, only a small set of large contractors can apply to these grants and even these companies must subcontract aspects of the project. The resulting project structure is unwieldy and prone to waste.[27] As one official interviewed for this project noted, the government has "processes that are built for the industrial age, not the information age."[28] In an era where technology cycles are five years (or less), the speed of big data innovation is critical. For example, a specific concern is maintaining up-to-date on-demand computing power (e.g. cloud services), which make it possible to store and analyze the most resource intensive analytics tasks.[29]

There are recent and positive signs of increasing big data adoption in CBP. In addition to the examples discussed above there are other efforts that are just now beginning. In 2018, the Border Patrol set up its own data science division.[30] Also, CBP has fielded an innovation team (INVNT) charged with technology transfer initiatives. The INVNT team was critical in the adoption and transition of technologies like the Android Team Awareness Kit (ATAK), which is a software application that helps users understand the position of other users as well as threats.[31] ATAK has the capability to import and display data from a variety of collection streams, such as unmanned aerial vehicles and other sensor platforms. Currently, there are 5,000 to 7,000 ATAK devices supporting a variety of roles, including the Border Patrol's Tactical Team (BORTAC).[32] These efforts are a good start to address the big data problem but additional steps are needed.

Recommendations

The benefits of increased big data applications at CBP are significant, and range from increased economic efficiency at ports of entry to improved allocation of enforcement resources along the border. To this end, CBP needs to develop a strategic framework for change that builds mechanisms to create and sustain a data-driven culture supported by the best available technology.

Recommendation 1: CBP needs to develop a big data capability roadmap aligned with the Federal Data Strategy.

Data governance and procedures are at the core of information sharing and collaboration in border security. Addressing these issues requires an evaluation of CBP's current needs and capabilities as well as steps moving forward. In particular, CBP needs to develop a big data capability roadmap that aligns with the Federal Data Strategy, a government-wide plan set in motion by President Trump's Management Agenda and which is expected to continue under the Biden Administration.[33] The plan lays out 20 action steps that provide "a common set of data principles and best practices in implementing data innovations that drive more value for the public."[34] For example, action step 1 requires organizations to identify their data needs and action step 2 to create a data governance body. In addition to these steps, the strategy document will outline observable milestones of a progression towards a data culture as well as developing procedures for better securing CBP's data. The subsequent recommendations presented here are elaborations of steps that can be taken as a part of a wider CBP data strategy.

Recommendation 2: CBP needs to define the mission and goals for a dedicated interdisciplinary team with the primary purpose to support CBP becoming more data-driven.

Developing a data culture is a necessary prerequisite for organizations to optimally use big data. To this end, CBP needs to create an interdisciplinary team that blends a combination of diverse individuals with data science skills—and just as importantly, operational experience—to develop the foundational principles of CBP's data-driven culture. This effort could complement or link to ongoing efforts, such as the Border Patrol's Data Science Team. An important member—and potential leader—of this team is a Chief Data Officer (CDO). CDOs are leaders who articulate overall data strategy and try to cultivate a data culture in an organization. In short, they serve as an organizational focal point to improve big data adoption. Several agencies have appointed a CDO, such as Department of Health and Human Services and Department Transportation. In addition to the CDO, the team should include some PhD-level data scientists to work on high-level problems, such as integrating information streams across the organization to surface new insights as well as individuals from across the organization with operational experience.[35]

CBP leadership also needs to take an important role in driving the shift towards a big data culture. Leadership in this area can take a variety of forms, such as positive messaging around data-driven decision making, and requiring education and human capital development programs like those described in recommendations 3 and 4.

Recommendation 3: CBP needs to invest in its workforce with a data science reskilling program and open new data science positions.

As noted earlier, the lack of data science human capital is one significant factor inhibiting the optimal use of big data in border security operations. Fortunately, it is not necessary for all employees to have a high level of knowledge in data science. Instead organizations require a core group of data science experts who can apply their knowledge to tasks, such as applying machine learning algorithms to large datasets. While efforts to recruit data scientists from outside CBP's ranks should continue, a more efficient strategy is to look within the organization. Initiatives, such as the Federal Cyber Reskilling program, can serve as a model for these efforts within CBP.

The Federal Cyber Reskilling program helps agencies identify current federal employees with the skills for cyber security positions but who are in non-cyber security roles. Selected employees attend courses and sharpen their skills over several weeks, and if successful, receive a certification.[36] A similar program for data science implemented in CBP will significantly boost the amount of human capital. It can tap into the vast and unused talent in CBP's large workforce. From our research for this article, we came across numerous cases of employees who have taught themselves data science skills and wanted to find more ways to innovate on the job.

Recommendation 4: CBP needs to develop and mandate basic data literacy training content for its employees.

While the vast majority of workers do not need to create or even use algorithms, most employees need basic knowledge to use big data. Google's chief decision scientist sums up this need well: it is not necessary to know how to build a microwave in order to use one effectively.[37] What personnel require is "data literacy," "the ability to read, write, and communicate data in context."[38] These skills are necessary because a growing number of CBP personnel are required to interpret ambiguous data on the job. For example, personnel are asked to assess the danger posed by travelers based on uncertain information, such as social media data.[39] To build data literacy, CBP should develop and mandate training for most of its employees. This training should begin when an employee begins work and continue periodically throughout their career to reflect technological advancements and tools.

Recommendation 5: CBP and DHS needs to facilitate private-sector technology transfer and improve the transition of commercial software on government systems.

Even if an organization develops a data culture and human capital, it cannot leverage big data without keeping up with the latest technological advancements. To address this issue, CBP should invest resources in technology transfer. As an immediate step, CBP needs to invest greater resources in its innovation team. As discussed above, this small group has had success transitioning private-sector innovations to the field and working between operational components of CBP. Another necessary task is for DHS to develop an external partner to help transition private-sector technology to the government.[40] Such "technology accelerators" provide the government an opportunity to interface with the private sector. For example, the CIA owns its own venture capital firm, which is funded but largely independent of that organization.[41] These organizations are important because they help speed up innovation by bringing in new companies and identify dual use technologies. A related issue is the need to transition commercial software to government systems efficiently. Guidance from Recommendation 1 above is critical in setting consistent standards for migrating software and resources to government systems. This point is especially important to prevent future breaches and safeguard data.

About the Authors

Stephen Coulthart is an assistant professor at SUNY Albany in the College of Emergency Preparedness, Homeland Security, and Cybersecurity. His research has been published in *Intelligence and National Security*, *International Affairs*, and the *Journal of Conflict Resolution*, among others. He is the lead editor of *Researching National Security Intelligence: Multidisciplinary Approaches* (Georgetown University Press). He is also a fellow with the Truman National Security Project.

Ryan Riccucci has held leadership positions in the U.S. Border Patrol overseeing intelligence and border security operations on both the southern and northern border. His career includes two headquarters assignments in Washington D.C., working with stakeholders across the homeland security enterprise on research, development, testing, and evaluation of innovative technology. Ryan is currently Deputy Patrol Agent In Charge (DPAIC) of a large station on the southern border and is an Adjunct Professor of Practice at the University of Arizona College of Applied Science & Technology.

Acknowledgements

The research for this article was made possible by a grant from Texas National Security Network's Excellence Fund. The authors also wish to thank the five anonymous reviewers for their invaluable feedback and suggestions.

Disclaimer

The views expressed are those of the authors and do not reflect the official policy or position of Customs and Border Protection (CBP) or the U.S. Government. CBP cannot attest to the substantive or technical accuracy of the information.

Notes

[1] Alan Bersin, “Lines and Flows: The Beginning and End of Borders,” *Brooklyn Journal of International Law* 37, no. 2 (2012), 389-406.

[2] Doug Laney, “3D data management: Controlling Data Volume, Velocity and Variety,” *META Group Research Note* 6, no. 70 (2001), 1.

[3] Austen D. Givens, Nathan E. Busch, and Alan D. Bersin, “Toward a Smarter Defense: Big Data and Risk Assessments in Homeland Security,” in *Homeland Security: An Introduction* (New York, NY: Oxford University Press), manuscript in preparation.

[4] Bram Klievink, Bart-Jan Romijn, Scott Cunningham, and Hans de Bruijn, “Big Data in the Public Sector: Uncertainties and Readiness,” *Information Systems Frontiers* 19, no. 2 (2017), 267-83.

[5] Givens et al., “Toward a Smarter Defense: Big Data and Risk Assessments in Homeland Security.”

[6] James Cullum, “Faced with Multiple Threats, CBP Initiates Data Sciences Division to Aid Physical, Cyber Challenges,” *Homeland Security Today*, November 8, 2018.

[7] Alan Bersin and Chappell Lawson, “Homeland Security and Transnational Crime,” in *Beyond 9/11*, eds. Chappell Lawson, Alan Bersin, and Juliette Kayyem (Cambridge, MA: MIT Press, 2020), 267-83.

[8] Interview with author, December 12, 2019.

[9] Department of Homeland Security, “Border Enforcement Analytics Program Apex Infographic,” https://www.dhs.gov/science-and-technology/beap-apex-infographic.

[10] Interview with author, December 4, 2019.

[11] Cullum, “Faced with Multiple Threats.”

[12] Francis X. Taylor, “DHS’ Big Data Integration Challenge,” *The Cipher Brief*, August 8, 2018, https://www.thecipherbrief.com/column_article/dhs-big-data-integration-challenge.

[13] Stevan Bunnell, “Increasing Security while Protecting Privacy,” in *Beyond 9/11*, eds. Chappell Lawson, Alan Bersin, and Juliette Kayyem (Cambridge, MA: MIT Press, 2020), 262-63.

[14] Letter from Hal Abelson et al. to Elaine C. Duke, November 16, 2017, available at https://www.brennancenter.org/sites/default/files/Technology%20Experts%20Letter%20to%20DHS%20Opposing%20the%20Extreme%20Vetting%20Initiative%20-%20 11.15.17.pdf.

[15] Faiza Patel, Rachel Levinson-Waldman, Sophia DenUyl, and Raya Koreh, *Social Media Monitoring: How the Department of Homeland Security Uses Digital Data in the Name of National Security*, (New York: Brennan Center, 2020), https://www.brennancenter.org/sites/default/files/2020-03/SocialMediaMonitoring.pdf.

[16] Office of the Inspector General, *Review of CBP’s Major Cybersecurity Incident during a 2019 Biometric Pilot*, OIG-20-71 (Washington, DC: Department of Homeland Security, September 21, 2020).

[17] Office of the Inspector General, *CBP Has Improved Southwest Border Technology, but Significant Challenges Remain*, OIG-21-21 (Washington, DC: Department of Homeland Security, February 23, 2021).

[18] Interview with author, November 26, 2019.

[19] Kevin C. Desouza and Jacob Benoy, “Big data in the public sector: Lessons for practitioners and scholars,” *Administration & Society* 49, no. 7 (2017), 1043-64.

[20] Interview with author, May 16, 2019.

[21] Ibid.

22 Kylie Bielby, “OIG: Obsolete Technology Compromises Southern Border Security,” *Homeland Security Today*, February 27, 2021.

23 For example, see: Sydney Freedberg, “Culture, Not Tech, Is Obstacle To JADC2: JAIC,” *Real Clear Defense*, February 11, 2021.

24 Chandana Gopal and Dan Vesset, “Why You Should Care About Data Culture,” IDC Research, April 2020, https://www.tableau.com/sites/default/files/pages/idc_infobrief_data_culture.pdf.

25 Interview with author, December 4, 2019.

26 Nate Bruggeman and Ben Rohrbaugh write about these issues in depth in Homeland Security Paper #5.

27 Interview with author, December 20, 2018. For example, a 2006 project, the Secure Border Initiative Network (SBI Net) was intended to create a surveillance system across the U.S.-Mexico border. The project stalled with only 53 miles completed at a cost of approximately a billion dollars. A 2010 OIG report concluded the project suffered from inadequate oversight of the project and contractors’ progress.

28 Interview with author, November 20, 2020.

29 Mark Rockwell, “CBP Expects Cloud Management Contract Next Year,” *The Business of Federal Technology*, August 13, 2020.

30 Cullum, “Faced with Multiple Threats.”

31 DHS Science and Technology Directorate, “Team Awareness Kit: Tactical Situational Awareness Solution,” March. 2020, https://www.dhs.gov/publication/team-awareness-kit-fact-sheet.

32 Interview with author, November 17, 2020; DHS Science and Technology Directorate, “ATAK: Android Team Awareness Kit,” YouTube Video, November 17, 2017, https://www.youtube.com/watch?v=HdNy0YBfEvU.

33 “Federal Data Management: An Agenda for the Administration’s First 100 Days,” *FedScoop Radio*, January 27, 2021.

34 Federal Data Strategy, “2020 Action Plan,” accessed December 9, 2020, https://strategy.data.gov/action-plan/.

35 For more on CDO best practices, see: Jane M. Wiseman, “Data-Driven Government: The Role of Chief Data Officers” (Washington, DC: IBM Center for the Business of Government, 2018), available at https://www.innovations.harvard.edu/data-driven-government-role-chief-data-officers.

36 Jory Heckman, “Federal Cyber Reskilling Academy to retrain federal employees as cyber defense analysts,” *Federal News Network*, November 30, 2018.

37 Cassie Kozyrkov, “Why Businesses Fail at Machine Learning,” accessed December 9, 2020, https://hackernoon.com/why-businesses-fail-at-machine-learning-fbff41c4d5db.

38 Kasey Panetta, “A Data and Analytics Leader’s Guide to Data Literacy,” accessed December 9, 2020, https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy/.

39 Patel et al. *Social Media Monitoring*, 2020.

40 Nate Bruggeman and Ben Rohrbaugh, “Closing Critical Gaps that Hinder Homeland Security Technology Innovation,” *Homeland Security Policy Paper* 5 (Harvard Kennedy School, Belfer Center for Science and International Affairs, April 2020).

41 In-Q-Tel, “About IQT,” accessed December 9, 2020, https://www.iqt.org/about-iqt/.