Journal of Strategic Security

Journal of Strategic Security

Volume 9
Number 2 *Summer 2016*

Article 6

Big Data and Intelligence: Applications, Human Capital, and Education

Michael Landon-Murray
University of Colorado, Colorado Springs, mlandonm@uccs.edu

Follow this and additional works at: https://digitalcommons.usf.edu/jss

pp. 92-121

Recommended Citation
Landon-Murray, Michael. "Big Data and Intelligence: Applications, Human Capital, and Education." *Journal of Strategic Security* 9, no. 2 (2016) : 92-121.
DOI: http://dx.doi.org/10.5038/1944-0472.9.2.1514
Available at: https://digitalcommons.usf.edu/jss/vol9/iss2/6

This Article is brought to you for free and open access by the Open Access Journals at Digital Commons @ University of South Florida. It has been accepted for inclusion in Journal of Strategic Security by an authorized editor of Digital Commons @ University of South Florida. For more information, please contact scholarcommons@usf.edu.

Big Data and Intelligence: Applications, Human Capital, and Education

Abstract

The potential for big data to contribute to the US intelligence mission goes beyond bulk collection, social media and counterterrorism. Applications will speak to a range of issues of major concern to intelligence agencies, from military operations to climate change to cyber security. There are challenges too: procurement lags, data stovepiping, separating signal from noise, sources and methods, a range of normative issues, and central to managing these challenges, human capital. These potential applications and challenges are discussed and a closer look at what data scientists do in the Intelligence Community (IC) is offered. Effectively filling the ranks of the IC's data science workforce will depend on the provision of well-trained data scientists from the higher education system. Program offerings at America's top fifty universities will thus be surveyed (just a few years ago there were reportedly no degrees in data science). One Master's program that has melded data science with intelligence is examined as well as a university big data research center focused on security and intelligence. This discussion goes a long way to clarify the prospective uses of data science in intelligence while probing perhaps the key challenge to optimal application of big data in the IC.

Acknowledgements

I would like to thank Stephen Coulthart for his generous input while developing this project. I'm also deeply appreciative of the feedback from the two anonymous reviewers.

This article is available in Journal of Strategic Security: https://digitalcommons.usf.edu/jss/vol9/iss2/6

Introduction

Big data and data science seem to have taken us by storm, while precise meanings, concepts, and applications continue to unfold. The Harvard Business Review has deemed data scientist the "Sexiest Job of the 21st Century,"[1] and proponents use quite bombastic rhetoric to highlight the benefits and opportunities afforded to any number of sectors. This is certainly matched with skepticism and apprehension connected to normative as well as practical issues. Big data and data science will no doubt bring important uses and benefits to people, commerce, government, and society, but serious issues of equity, privacy, epistemology, and methodology, among others, exist and need attention at this early juncture.

This article will be one of the first detailed looks at data science and big data in the context of intelligence, considering applications, possibilities, limits, and challenges of these emerging fields. The analysis extends in more detail to perhaps the most immediate challenge–that of human capital and education–through a multi-pronged approach. This includes looking at the actual job of data scientist in the Intelligence Community (IC). Just four years ago, academic programs in data science were found to be very rare if not virtually nonexistent, a supply challenge for sure![2] This is not only a critical dimension in a short-term manpower sense, but one that will require attention if the full and optimal use of big data and data science is to be realized in the IC and the central issues and limits successfully addressed.

To investigate data science and big data in the higher education system, this article maps out academic programs in big data, data science, and the tightly related field of analytics in America's top fifty national universities, as ranked by the US News and World Report. A case example of a program that expressly melds data science and intelligence analysis, located at Mercyhurst University, is also presented. The program at Mercyhurst equips its data science students with domain expertise in intelligence studies and analysis. This combination will be critical for the IC, as will be shown, melding the technical with the substantive. The IC is also likely to be at a comparative disadvantage *vis-a-vis* its private sector competitors in attracting data scientists. Companies will by-and-large be able to compensate their data scientists–which are and will continue to be in high demand–at higher levels.

[1] Thomas H. Davenport and D.J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review* 90:10 (October 2012): 70-76.
[2] Ibid.

Thus, the IC might be well served by an early and specialized pipeline to future big data practitioners, and Mercyhurst offers us insight into a first step in that direction. Reviewing this program will also help keep the intelligence education literature up-to-date on newly emerging and unique programs.

Institutions of higher education are also a source of research and technology for all kinds of industries, including intelligence. With that in mind, this article will also examine a university-based research center that focuses on using big data and data science to further security and intelligence missions. This will also further clarify what kinds of big data applications the IC might utilize, and such centers can aid in providing needed technologies to support IC data science and scientists. This case example also demonstrates the way that the federal government is supporting and partnering with academia to help build its own capacities in these areas. Of course, it is just one example, and action has been taken across the federal government to advance and take advantage of big data, including the White House's Big Data Initiative for Research and Development, the Defense Department's Data to Decisions program, a number of projects at the Intelligence Advanced Research Projects Activity (IARPA), and the National Science Foundation's Regional Big Data Innovation Hubs (also university partnerships).

Before looking at the role of higher education, this article begins by considering what big data, data science, and analytics are, and what benefits, issues, and problems can accompany the phenomenon of big data. Many of these dynamics will have important implications for intelligence applications, though use in intelligence organizations will carry unique challenges which are treated separately in this article. The parameters and connections associated with big data, data science, and analytics are contested and evolving, to be sure, and this article can only establish foundations and introduce key applications and conversations. This discussion of issues and challenges will then extend to the realm of intelligence, where specific applications will also be reviewed and data scientist job descriptions at the Central Intelligence Agency (CIA) and National Security Agency (NSA) mapped and discussed. As one expert has said, "data science is what data scientists do."[3] With that in mind, one of the best ways to investigate data science human capital dimensions–and the very concepts of what big data and data science are in the IC–is simply to look at the job of data scientists in the IC. The article concludes with a discussion of key observations, including consideration of what precepts, skills, and characteristics of effective data

[3] Schutt, Rachel, "Educating the Next Generation of Data Scientists," *DataEDGE 2013*, UC Berkeley School of Information, Berkeley, California, May 31, 2013.

DOI: http://dx.doi.org/10.5038/1944-0472.9.2.1514

scientists can inform the approaches and ethic of intelligence analysis more generally.

Big Data, Data Science and Analytics: Definitions, Applications and Issues

The definitions and concepts of big data, data science, and data analytics are emerging, being contested, and evolving, and that will continue. Big data is being used across many sectors, spanning business and marketing, health, policy, and even politics. In the public and political sectors, police agencies are using big data to anticipate criminality and political campaigns are using massive data sets and analytics to target and persuade voters. The presidential campaigns of Barack Obama have relied heavily on big data, data analytics and data scientists, parsing individual potential voters along 80 different dimensions to identify preferences and attributes, in a sense enabling "data-driven pandering."[4] The McKinsey Global Institute estimates that by 2018 the U.S. economy will have a shortage of between 140,000 and 190,000 individuals with the requisite deep technical skills for big data analysis.[5]

Rachel Schutt, currently Chief Data Scientist at News Corps and adjunct faculty at Columbia University, has pointed not only to unclear definitions in the media, but important underlying questions of how big does data need to be to be big data, if data science is simply the science of big data, and moreover if data science is more than statistics.[6] As Schutt notes, "Statisticians already feel that they are studying and working on the 'Science of Data.'"[7] John DeLong, Director of Compliance at the NSA, has also noted the still contested nature of big data.[8]

[4] L. Gordon Crovitz, "Obama's 'Big Data' Victory," *Wall Street Journal*, November 18, 2012, available at: *http://www.wsj.com/articles/SB10001424127887323353204578126671124151266.*
[5] McKinsey Global Institute, *Big Data: The Next Frontier for Innovation, Competition and Productivity*, June, 2011, available at*: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.*
[6] Schutt, "Educating the Next Generation of Data Scientists"; O'Neil, Cathy and Rachel Schutt, *Doing Data Science: Straight Talk from the Frontline* (Sebastopol, CA: O'Reilly Media, 2013), 2.
[7] O'Neil and Schutt, *Doing Data Science.*
[8] DeLong, John, "National Security Implications of 'Big Data' Surveillance," *Dartmouth College Surveillance in the Age of Big Data Spring Speakers Series*, Steele Hall, Dartmouth College, Dartmouth, New Hampshire, May 14, 2014.

Standard conceptions of big data focus on various "V's"–the volume, velocity, value, veracity, and variety of what are petabytes of data. Further, big data is often captured, integrated, and used in real-time; it can come in structured or unstructured forms, is meant to be supremely comprehensive but also very precise, and can be easily combined with other data sources to grow rapidly.[9] Thus, a core thrust of big data is the idea that we can move beyond sampling, essentially having all the N's–a complete population. Elizabeth D. Liddy of Syracuse University has described big data simply "as a very large collection set of either textual or numeric data and, increasingly, image data…The clear goal is to reuse it."[10] Lastly, Danah Boyd and Kate Crawford have captured the concept of big data this way:

> "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets…We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of:
>
> 1. *Technology*: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
> 2. *Analysis*: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
> 3. *Mythology*: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy."[11]

Without question, the digitization of information has propelled big data forward, but this also has been joined by the "datafication" of behaviors, conditions and characteristics of many more things than might have been hoped for even twenty years ago–including posterior recognition systems as anti-theft devices![12]

[9] Rob Kitchin, "Big Data and Human Geography: Opportunities, Challenges and Risks," *Dialogues in Human Geography* 3:3 (November 2013): 262-267.

[10] Edd Dumbill and Elizabeth D. Liddy, Jeffrey Stanton, Kate Mueller, and Shelly Farnham, "Educating the Next Generation of Data Scientists," *Big Data* 1:1 (March 2013), 21.

[11] Danah Boyd and Kate Crawford, "Critical Questions for Big Data," *Information, Communication and Society* 15:5 (June 2012): 662-679.

[12] Kenneth Cukier and Viktor Mayer-Schonberger, "The Rise of Big Data," *Foreign Affairs* 92:3 (May/June 2013), 27-40.

https://digitalcommons.usf.edu/jss/vol9/iss2/6
DOI: http://dx.doi.org/10.5038/1944-0472.9.2.1514

Just as important as the datafication of so much of society is the emergence of technologies and programs that are capable of processing and analyzing the immense volume and diversity of data now available. Such technologies have proliferated in recent years. Big data allows for data mining analytics, creating models and algorithms to help cull out behaviors, patterns, relationships, and predictions. New tools and programs, such as Hadoop and Cosmos, allow data scientists to organize, process, and analyze massive amounts of data. Such systems and tools will continue to emerge and evolve, and data scientists will need to stay abreast of such developments and the opportunities they afford. The shrinking costs of big data processing capacities are certainly a considerable factor in their expansion within and across sectors.[13]

The hype and hope around big data and data science in the last several years has been of tremendous proportions, at least in some quarters. The optimism around what data science and big data can accomplish have been viewed as exaggerated by others, including those in the field. The Harvard Business Review called data scientist the "Sexist Job of the 21st Century" and equally dramatic rhetoric on the benefits of big data abound. Christopher Anderson, writing in Wired, has suggested that big data transcends longstanding approaches to knowledge advancement and makes context and causation obsolete, leaving correlation.[14] Critics would contend that numbers do not "speak for themselves," that context matters, and that the "why" of phenomenon remain very important–it surely is possible that patterns or connections will be made where they do not actually exist.[15] David J. Leinweber presents one such false connection between fluctuations in the S&P 500 index and butter production in Bangladesh.[16] Further, sampling remains a very real issue, and where and how you get your data–and thus what data you get–matters, an issue we will return to below.[17]

[13] Kevin S. Bankston and Ashkan Soltani, "Tiny Constables and the Cost of Surveillance: Making Cents out of the United States Versus Jones," *The Yale Law Journal* 123 (January 9, 2014), available at: *http://www.yalelawjournal.org/forum/tiny-constables-and-the-cost-of-surveillance-making-cents-out-of-united-states-v-jones*.

[14] Christopher Anderson, "The End of Theory: The Data Deluge makes the Scientific Method Obsolete," *Wired*, June 23, 2008, available at: *http://www.wired.com/2008/06/pb-theory/*.

[15] Boyd and Crawford, "Critical Questions for Big Data."

[16] David J. Leinweber, "Stupid Data Miner Tricks: Overfitting the S&P 500," *Journal of Investing*, 16:1 (Spring 2007): 15-22.

[17] O'Neil and Schutt, *Doing Data Science*; Crawford, Kate, "The Raw and the Cooked: The Mythologies of Big Data," *DataEDGE 2013*, UC Berkeley School of Information Science, Berkeley, CA, May 30, 2013; Boyd and Crawford, "Critical Questions for Big Data."

Rachel Schutt has raised concerns about "charlatans," "rock stars," and "frauds" who are concerned with self-promotion rather than "the underlying ideas."[18] Mark M. Lowenthal worries that the U.S. Intelligence Community is also susceptible to fads, writing, "Most likely…big data will not get at the questions that most bedevil policymakers and analysts–intentions."[19] In the sorts of big data collected, there certainly is a bias in favor of indicators that can be more discretely and easily measured (for example, online purchasing decisions) and counter to normative values, emotions, beliefs and interactions.[20] This could be considered a critique of the "Big Data hubris"—an outlook or expectation that big data can "substitute for rather than supplement conventional modes of analysis."[21] Demonstrating the efficacy of both the data and the analytic methods will be critical.[22] And as Rob Kitchin has written, "Whilst there has been much recent progress in devising new data analytics that can make sense of massive data sets, new forms of data science are in their infancy."[23] Danah Boyd and Kate Crawford add, critically,

> "…the specialized tools of Big Data also have their own inbuilt limitations and restrictions. For example, Twitter and Facebook are examples of Big Data sources that offer very poor archiving and search functions. Consequently, researchers are much more likely to focus on something in the present or immediate past…because of the sheer difficulty or impossibility of accessing older data."[24]

Kate Crawford also points to what she labels six myths about big data, the first being that big data is something brand new.[25] Rather, she suggests that its pervasiveness and availability is new, and she points to industries with big data in their past, including the NSA and the Census Bureau. Schutt similarly observes, "From the way the media describes it, machine learning algorithms were just invented last week and data was never 'big' until Google came along. This is simply not the case."[26] Crawford also worries that big data might give the veneer of objectivity, and be used to better the lot of the privileged more

[18] Schutt, "Educating the Next Generation of Data Scientists."
[19] Mark M. Lowenthal, Intelligence Education: *Quo Vadimus*?" *American Intelligence Journal* 31:2 (2013): 7-11.
[20] Kitchin, "Big Data and Human Geography."
[21] David Lyon, "Surveillance, Snowden and Big Data: Capacities, Consequences and Critique," *Big Data & Society* 1:2 (July-December 2014): 1-13.
[22] Ibid.
[23] Kitchin, "Big Data and Human Geography," 264.
[24] Boyd and Crawford, "Critical Questions for Big Data."
[25] Crawford, "The Raw and the Cooked."
[26] O'Neil and Schutt, *Doing Data Science*, 2.

so than the poor.[27] She recalled the post-Sandy story painted by New York City Twitter feeds, one that captured the experience of those with Twitter accounts, which is not at all representative of the broader population and its experience. Discriminatory uses of big data can include the use of models to predict sensitive characteristics (such as income, race, political leanings and sexuality) and determine what sorts of information and offers different classifications of people are likely to receive. She additionally highlights problems with the notion that big data will naturally enhance policy, giving the example of making cities "smart," which can produce policies that are counter to basic social equity—that is, they can tend to be of help mostly to the more affluent. As she notes, big data is only as good as the people using its tools, and as we will see below, data science is as much an art as a science (a reality that adds to the subjectivity of big data).

Related to discrimination is the idea that you can "opt out" of having your signatures exempted from becoming part of big data. Services that allow this opt out cost money, and again will favor the wealthier rather than lower income households.[28] Recently, we have seen that parents sometimes do not have the choice of opting their children out of Google's collection and tracking of student data.[29] Further, the combination of data sets can also compromise the privacy and identity of individuals.[30] Thus, anonymity issues are very real in the domain of big data, and can present regulatory challenges, including in the areas of health information and even Smart Grid data, which can reveal very specific personal and home behaviors.

What is a Data Scientist, Analyst?

According to one of the individuals who popularly coined the term 'data scientist' (after deciding against 'data artist'), data scientists are "high-ranking professionals with the training and curiosity to make discoveries in the world of big data."[31] By "swimming" in this sea of data, data scientists access, explore, integrate, analyze and assess disparate and often unstructured sources of big data, creating mash-ups in the now popular parlance. Hilary Mason and Christopher Wiggins have described the data science process as

[27] Crawford, "The Raw and the Cooked."
[28] Ibid.
[29] Andrew Peterson, "Google is Tracking Students as it Sells More Products to Schools, Privacy Advocates Warn," *The Washington Post*, December 28, 2015, available at: *https://www.washingtonpost.com/news/the-switch/wp/2015/12/28/google-is-tracking-students-as-it-sells-more-products-to-schools-privacy-advocates-warn/.*
[30] Alessandro Acquisti and Ralph Gross, "Predicting Social Security Numbers from Public Data," *Proceedings of the National Academy of Sciences* 106:27 (2009): 10975-10980.
[31] Davenport and Patil, "Data Scientist."

OSEMN (sounds like "possum"): obtaining, scrubbing, exploring, modeling and interpreting data.[32] Jeffrey Stanton of Syracuse University's School of Information Studies divides data science into analytics, infrastructure and technology, and curation.[33]

Perhaps more succinctly, and a bit tongue-in-cheek, Josh Wills, Director of Data Science at Cloudera, has said that a data scientist is a "person who is better at statistics than any software engineer and better at software engineering than any statistician."[34] Even more succinctly, Schutt concludes that data science "is what data scientists do,"[35] suggesting its diverse and fluid uses, which are broad enough for data science to be "the study of the space of problems that can be solved with data."[36] She endorses a team-based approach to data science, bringing together all the different skill and knowledge sets that specific complex, interdisciplinary big data questions and challenges require. Schutt concludes that one of the most critical challenges and tasks is to "Map the space of problems that can be solved with data to the teams of people whose skill sets could construct a solution."[37]

The differences between data science and data analytics are certainly "blurry,"[38] using many of the same tools for the same ends to inform decision making,[39] something that is reflected also in the design and nature of data science and analytics academic programs. As we will also see below, U.S. intelligence agencies are in part looking for those with data analytics degrees to fill their data scientist ranks. Ryan Swanstrom, the Microsoft data scientist who maintains Data Science 101, has suggested that data scientists will spend more time designing algorithms and data analysts will spend more time applying existing algorithms, and thus may not be as technically trained as their data scientist counterparts. However, it increasingly seems that special analytics degrees will replace the previous degrees analytics specialists may have completed. Swanstrom's fantastic list of "Colleges with Data Science

[32] Hilary Mason and Christopher Wiggins, "A Taxonomy of Data Science," *Dataists.com*, September 25, 2010, available at: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/.
[33] Dumbill et al., "Educating the Next Generation of Data Scientists," 21.
[34] Data Science 101, "What is a Data Scientist?" October 16, 2012, available at: *http://101.datascience.community/tag/software-engineer/*.
[35] Schutt, "Educating the Next Generation of Data Scientists."
[36] Ibid.
[37] Ibid.
[38] Ryan Swanstrom, "Analytics vs. Data Science," *Data Science 101*, October 24, 2015, available at: *http://101.datascience.community/2015/10/24/analytics-vs-data-science/*.
[39] Phil Simon, "Comparing Analytics and Data Science," *American University*, October 17, 2015, available at: *https://903ink.onlinebusiness.american.edu/blog/comparing-analytics-data-science/*.

Degrees" incorporates big data, data science and analytics degrees, accordingly.[40]

In her introductory—even exploratory class—on data science at Columbia University, Schutt offered seven disciplinary "buckets" that focus on the harder component skill sets of the field: data visualization, machine learning (automated, advancing algorithms), mathematics, statistics, computer science, communication, and domain expertise.[41] Jeffrey Stanton stresses also that data scientists will need to have specific domain knowledge specific to the field they work or study in to have the most success.[42]

Additionally, data scientists must be good at communicating and storytelling, verbally and visually, a set of attributes that go beyond science. This suggests there is a far degree of art in data science, a dichotomy—or combination—that reflects similar conversations about whether intelligence analysis is an art or a science. Rachel Schutt has pointed to the "habits of mind" that good data scientists have, noting storytelling, data intuition, and curiosity, as well as an ability to interpret.[43] Thus, it is yet again evident that big data does not afford an objectivity that yields unadulterated outputs, but entails a creative process that certainly—and healthily—allows for human subjectivity to enter the picture.

Big Data and Intelligence: Applications, Opportunities and Limits

The academic literature on big data and intelligence analysis and operations has remained largely abstract, with more narrow focus coming in the areas of bulk data collection and counterterrorism (and the attendant civil liberties and privacy issues). The most detailed assessments take on the Snowden revelations, social media and counterterrorism more generally. Brant C. Reilly adroitly wonders if the bulk collection programs of the NSA have absorbed or redirected attention away from other potential uses and benefits for the IC.[44]

[40] Ryan Swanstrom, "Colleges with Data Science Degrees," *Data Science 101*, February, 2016, available at: *http://101.datascience.community/2012/04/09/colleges-with-data-science-degrees/*.
[41] Schutt, "Educating the Next Generation of Data Scientists."
[42] Dumbill et al., "Educating the Next Generation of Data Scientists," 22.
[43] Schutt, "Educating the Next Generation of Data Scientists."
[44] Brant C. Reilly, "Doing More with More: The Efficacy of Big Data in the Intelligence Community," *American Intelligence Journal* 32:1 (2015): 18-24.

First and foremost, it is necessary to lay out exactly what sorts of intelligence questions and issues big data can help address, lists that will continually grow. As noted, Mark M. Lowenthal has said that some of the most difficult intelligence challenges—such as discerning the intentions of foreign leaders—will not be clarified by big data.[45] There may simply be no big data on such matters. On its face, this may be a valid statement and sentiment (and this author tends to agree), but big data can certainly allow for the identification of previously unknown relationships. Even if the context and causality of those relationships remain unknown, it is possible that intentions—or perhaps more important, behavior—are correlated with currently unknown factors. This could be of huge help to analysts and policymakers, though does point to another issue. It will be a challenge to explain analytical insights premised on big data to policymakers, who will hopefully ask responsible, hard questions, and be skeptical of taking action on the basis of foggy relationships. Thresholds and policies will need to be established regarding the taking of various operational, investigative and policy steps in response to what will sometimes be uncertain correlations.

It will in part be up to creative data scientists, with broad and interdisciplinary support including other intelligence analysts, to cull out any such potential underlying relationships and convey them to policymakers. As Brant C. Reilly has noted, the extent of possibilities in this realm remains unknown, though his perspective seems to relegate concerns about privacy to perhaps a worrisome minimum, writing:

> "…leaders in the IC should lobby for policy that does not limit collection, but instead imposes limits on how the data are used, as this may appease privacy advocates. The limit of potential for big data applications is unknown; thus, barring collection could limit societal development."[46]

Kevjn Lim has written on the uses of big data for strategic intelligence analysis, but specifics in terms of uses and sources are mostly absent. He does discuss the use of social media as a source of sentiment analysis that can help intelligence agencies and policymakers anticipate longer-term trends and strategic change (such as the Arab Spring movement). This is a huge benefit and step forward, for sure. As Elizabeth D. Liddy has noted,

> "…there is just tons of data, way too much for anyone to be able to read and get a picture either of an organization, or of an individual…by

[45] Lowenthal, "Intelligence Education."
[46] Reilly, "Doing More with More."

> doing the natural language processing of it, particularly the sentiment analysis, the emotional side, you are able to provide insights and answers very quickly that would have taken analysts days and weeks of reading in order to come up with a similar picture."[47]

Bulk collection and data mining of telephonic metadata and computer transmissions (PRISM) are the two areas that have received the most attention in the media and public. These programs were implemented chiefly for the purposes of counterterrorism. The Snowden revelations led to public outcry and dialogue on the NSA's bulk collection programs as concerns about privacy and civil liberties came to the fore. Multiple studies, including an independent study commissioned by the White House and another by the New America Foundation, have found these programs to be of limited efficacy in preventing terrorist attacks.[48] Of course, larger questions of public policy and norms in a democratic society are key, and to a degree rest in the hands of the voting public. Even federal judges have differed on their constitutional interpretation of NSA bulk collection. However, late last year, at the behest of Congress and the President (by way of The USA Freedom Act), the NSA ended its bulk collection of American telephonic and internet metadata.

With the now pervasive use of drones in U.S. military and counterterrorism operations, including drones that can be continually sucking in video feed of huge swaths of area, the amount of data being collected exceeds the analytic capacity for that intelligence. While this is not a new problem in general, it is a new opportunity that brings challenges for sure. The Department of Defense is working on methods for coping with this—as Sean Fahey notes, this new intelligence source creates literally years and years of footage, and at an increasing rate from one year to the next.[49] A related issue with new "big" sources of intelligence that can support those in the battlefield is having the coverage and connectivity to warfighters and other operators on the ground. This of course also creates new and dense information flows that operators must learn and incorporate into their actions. More generally, Fahey reports that analytics specialized for defense and military operations remain

[47] Dumbill et al., "Educating the Next Generation of Data Scientists," 26.

[48] New American Foundation, *Do NSA's Bulk Surveillance Programs Stop Terrorists?* (Washington D.C.: New America Foundation, January 2014), available at: *https://www.newamerica.org/downloads/IS_NSA_surveillance.pdf*; The White House, *Liberty and Security in a Changing World: Report and Recommendations of the President's Review Group on Intelligence and Communications Technologies* (Washington, D.C.: The White House, December 2013), *available at: https://www.whitehouse.gov/sites/default/files/docs/2013-12-12_rg_final_report.pdf.*

[49] Fahey, Sean, "Big Data and Analytics for National Security," *Workshop on Algorithms for Modern Massive Data Sets*, Stanford University, Stanford, CA, July 10-13, 2012.

somewhat immature, partly a function of the secret nature of relevant information and needs.

Opportunities to support security and intelligence missions are also present in areas like port security (smuggling and trafficking), money laundering and cyber security.[50] Activity and transactions can be mined and combined to identify probable illicit activity, and again, it will be up to data scientists in this realm to identify what patterns seem to correlate with crimes and threats in these domains. Security and intelligence data scientists should be able to devise models that allow them to identify when these systems are being misused or abused, hopefully better separating the signal from the noise and "finding the needle in the haystack" more effectively—assuming it is known what the needle will look like.

Big data and data science are also creating pertinent opportunities in areas like climate change and pandemics. Severe weather events are proving among the most serious threats to homeland and national security, and big data and supercomputing models produce simulations that can help predict changes and impacts. Just as Google has shown some temporary success in anticipating pandemics,[51] data scientists at intelligence agencies may be able to predict various public health problems in a more rapid fashion. This could be especially important where other detection and response systems are inadequate or lagged in some way. As we've experienced in the last couple years, pandemics in other parts of the world can spread internationally and create public health emergencies in places all over the world. But, and as the Google example demonstrates, such analyses may have fleeting accuracy, and "If you have no idea what is behind a correlation, you have no idea what might cause that correlation to break down."[52]

There are without question myriad other issue areas and uses big data and data science will help intelligence and security organizations navigate. It may be that new models and algorithms give political and financial insights not previously possible. As will be seen in the section below on academia's role in big data, tools are also being devised to aid in the detection of weapons of mass destruction and human trafficking, among other problems and threats. Opportunities and applications across the spectrum of more traditional as

[50] Ibid.
[51] Cukier and Mayer-Schonberger, "The Rise of Big Data."
[52] Tim Harford, "Big Data: Are We Making a Big Mistake," *Financial Times*, March 28, 2014, available at: ***http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html***.

well as emerging security issues will likely avail themselves. And as new kinds of big data emerge, new uses will follow, and data scientists will certainly exercise agency in identifying or constructing new data and data sets.

Whatever the varied potential, there are additional big picture constraints that will challenge the full realization of big data benefits for U.S. intelligence. Somewhat obviously, governmental budgeting and procurement processes can create lags in both technology and use. Human capital and recruitment constraints in this area will continue to challenge the IC,[53] constraints that presumably should be reduced as more data science and analytics graduates are produced by our higher education system. John Delong, compliance chief at the NSA, has also noted challenges in public understanding of big data and issues of trust when it comes to big data programs in the Intelligence Community.[54] Thus, and if the IC is to use big data in a sustainable way, the public and their elected representatives will need to comprehend the role of big data in national and homeland security. If our society and democracy are to approach this era of booming information and potential intrusion in a way consistent with democratic accountability and norms, the broad contours of these new applications and technologies will need to be understand. Of course, the specific methods and uses will not be part of the public knowledge.

Another impediment to the optimized use of big data in the Intelligence Community is that access to data both domestically and internationally can—and should—have limits. Collection limits at home will reduce the amount of big data available to intelligence agencies, and while those limits will not constrain foreign collection to the same extent, some types of potentially useful data that have been shown to help model and predict various outcomes and behaviors in some contexts may simply not be captured, cleaned and made available. Like in other intelligence sources, denial and deception can also impede access to big data as potential targets learn and work to be evasive and even manipulative. Further, rare events—which can be connected to intentions—may not be associated with big data sources, at all or in ways that can be recognized, and thus analytic insights from data scientists will probably be minimal in these areas. And historical issues of "stovepiping" will surely present limits on the sources of big data that are available, at least for the foreseeable future.

Other issues also remain to be investigated, including for example how data science teams in the IC should be structured and then integrated with or

[53] Fahey, "Big Data and Analytics for National Security."
[54] DeLong, "National Security Implications of 'Big Data' Surveillance."

connected to other intelligence analysts and policymakers. Conversely, would data scientists best be distributed around agencies and analytic groups? Possibly both? Similarly, should data scientists provide analytic inputs to existing products and/or provide standalone assessments? How will other analysts react to this new class of intelligence practitioner, and what additional training might they need to best interact with data scientists? How can intelligence analysts both understand and vet the information sources and analysis received from their data science counterparts to confidently enhance intelligence products? Experimentation and evidenced-based practice will best resolve these issues and questions, helping arrive at the most seamless and value-adding practices and relationships.

Additionally, what sort of involvement can and should IC data scientists have with professional communities outside of the IC, beyond continuing education in data science and analytics? Will policymakers and managers feel confident in their own understanding of big data and data science to take policy or management action? It might also be worth investigating policymaker education and understanding in these areas.

The article now turns to an examination of what it means in practice to be a data scientist in the U.S. Intelligence Community, focusing on the CIA and NSA. Of course, actual data scientists will not be the only analysts in the IC with big data responsibilities, and as a former CIA Chief Technology Officer has noted, the IC is seeking big data applications and analytics that allow non-technical users to interact with big data via systems that allow relatively simple requests to produce complex and insightful outputs.[55] Educating students and future data scientists to be "comfortable with algorithms and data analysis as well as with social analysis and theory" is no doubt a challenge.[56]

Data Scientists at the Central Intelligence Agency (CIA)

If data science is best defined by what data scientists actually do, then defining data science in the context of intelligence suggests looking at what data scientists are expected to do in the Intelligence Community. In this section, data science at the CIA will be discussed, followed by data science at the NSA. Of course, looking at job listings provides us only the broad contours, but is probably our best option at this point, and surely it is in

[55] Hunt, Ira. "The CIA's 'Grand Challenges' with Big Data," *GigaOM Structure Data 2013*, New York City, March 20, 2013.
[56] Boyd and Crawford, "Critical Questions for Big Data," 674.

intelligence agencies' benefit to accurately represent the work of their (future) data scientists.

The very first lines of the CIA's data scientist job description read, "Do you have a passion for creating data-driven solutions to the world's most difficult problems? The CIA needs technically-savvy specialists to organize and interpret Big Data to inform U.S. decisionmakers, drive successful operations, and shape CIA technology and resource investments."[57] Not surprisingly, big data is put front and center in the work of data scientists at the Agency and across a range of supports. To fill these positions, the CIA is looking for graduates who have the following backgrounds:

- Data Analytics
- Computer Science
- Mathematics
- Statistics
- Economics
- Operations Research
- Computational Social Science
- Quantitative Finance
- Engineering
- Other data analysis fields

Entry-level requirements include a relevant Bachelor's degree, "and experience with applied quantitative research working with real world data, either through thesis research, internships, or work experience."[58] New hires can also enter at the "Developmental" or "Full Performance" levels, with increasing educational and experiential requirements. Those coming in at the Developmental level must have a Bachelor's or Master's degree and two to five years of experience in data science, or a closely related field. Applicants should "have demonstrated ability to successfully complete projects with large or incomplete data sets and be able to provide solutions."[59] At the Full Performance level, hires are required to have a Master's degree or equivalent work experience, as well as five or more years of critical experience. Applicants "should be an expert in their field and have demonstrated ability

[57] Central Intelligence Agency, "Careers & Internships," March 26, 2015, available at: https://www.cia.gov/careers/opportunities/science-technology/data-scientist.html.
[58] Central Intelligence Agency, "Careers & Internships."
[59] Ibid.

leading interdisciplinary teams throughout the full course of a project's life-cycle."[60]

Further, successful applicants "will have keen technical insight, creativity, initiative, and a curious mind."[61] Once candidates are selected and arrive, they will help "develop computational algorithms and statistical methods that find patterns and relationships in large volumes of data." The listing rightly boasts that CIA data scientists will have access to data sets not available to those outside of the intelligence world. With the products and findings produced, data scientists will need to be able to communicate in a clear, cogent fashion with a "lay audience" in the form of written and oral presentation.[62] Candidates are required to submit a short writing sample demonstrating their competencies and current research projects. It is also preferred that candidates have at least some experience and expertise in advanced statistical or computational programming and knowledge of modern databases, distributed computing systems, and the leading big data information management platforms.

The description of data scientist at CIA reflects many of the key components offered by data science scholars. This includes the need to be highly creative and curious, and the necessity that data scientists work in interdisciplinary teams rather than alone to be most effective. What is generally absent from the job description are the specific types of "accounts," issues and questions that data scientists are likely to be assigned—though the earlier part of this section provides at least some good indications.

Data Scientists at the National Security Agency (NSA)

At the NSA, data scientists are described as computer scientists

> "Who provide the link between the analytic, technical and development communities at NSA. They develop scalable analytic solutions to the workforce and apply them to NSA's largest data problems in order to support NSA's intelligence analysis missions."[63]

[60] Ibid.
[61] Ibid.
[62] Ibid.
[63] National Security Agency, "NSA Careers," February 2, 2016, available at: *https://www.nsa.gov/psp/applyonline/EMPLOYEE/HRMS/c/HRS_HRAM.HRS_CE.GBL?Page=HRS_CE_HM_PRE&Action=A&SiteId=1.*

DOI: http://dx.doi.org/10.5038/1944-0472.9.2.1514

Thus, NSA data scientists are a “hybrid of computer scientist and analyst,” bridging the highly technical with the analytical. Responsibilities of data scientists at NSA can look as follows:

- “Combine information about the structure, syntax, and processing of data with the functions of gathering, organizing, and manipulating datasets in order to synthesize responses to customer information needs.
- Apply scientific techniques to data evaluation, performing statistical inference and data mining.
- Document and present the data analysis and its conclusions for assessment by full-performance analysts, developers, and their managers.
- Gather requirements for recommended or needed improvements.
- Develop analytic plans, engineer supporting algorithms, design and implement solutions which execute analytic plans.”[64]

Entry-level and developmental data scientists may begin at the NSA with a high school degree and three years of relevant experience, an Associate’s degree and one year of experience, or a Bachelor’s degree and no experience. Degrees or minors must be in a STEM discipline. Relevant experience includes data mining, informatics, data science, programming, computational algorithms, information retrieval, statistical analysis, machine learning, artificial intelligence, software engineering and/or systems design and analysis. If the STEM Bachelor’s degree is in the social or behavioral sciences, graduates must have completed one or more classes in either research methods or math, such as applied statistics and calculus. If the STEM requirement is met through an academic minor or Associates degree, graduates must have taken at least three upper-level courses in applied statistics, calculus, quantitative or statistical methods, data mining, informatics, data science, programming, computational algorithms, information retrieval, databases or data structures, statistical analysis, machine learning, artificial intelligence, software engineering or systems design analysis. The completion of a data science certificate can be substituted for one year of relevant experience.

Data scientists can also begin employment at the NSA at the “Full Performance” level. Individuals can enter at this level with a high school diploma and four years of relevant experience, an Associate’s degree and two

[64] Ibid.

years of experience, a Bachelor's with one year of experience, or a Master's degree with no experience required. Course requirements for those who completed a social or behavioral science degree or an Associate's or minor in a STEM discipline are much the same as for the entry level data scientists, as is the relevant work experience. Candidates entering under this status can also apply completion of a data science certificate to replace a year of relevant experience. The above was taken from one of a small handful of data scientist openings at the NSA, and there is some variation in job descriptions, while education and experience requirements tend to be the same.

Higher Education, Big Data and Intelligence

A Note on U.S. Intelligence Education

Prior to 2001, Mercyhurst University was the only civilian school in the country to offer an intelligence studies degree, followed by a handful in 2001, and then dozens of concentrations, certificates and degrees after that. There is certainly contention around how these programs should be approached (as standalone programs or minors within degrees, for example).[65] The introduction of new programs has continued to the present, partly propelled by the U.S. Intelligence Community Centers for Academic Excellence (ICCAE) grant program. ICCAE, administered by the Defense Intelligence Agency, continues to support program development of various kinds in colleges and universities and to date has not been critically examined in the intelligence education literature.

There are now at least thirty undergraduate and graduate degrees in intelligence, spanning multiple spheres—cyber, analysis, intelligence studies, national security, business—and delivered in brick and mortar classrooms, online, and in hybrid fashion. William C. Spracher was the first to look closely at these programs, finding, among other things, that they speak well to IC Core Competencies.[66] Others have found such programs to be deficient in social science methods and models,[67] information processing and knowledge

[65] William C. Spracher, "National Security Intelligence Professional Education: A Map of U.S. Civilian University Programs and Competencies" (Doctoral Dissertation: The George Washington University, 2009); Noel Hendrickson, "Intelligence Analysis as an Academic Discipline: A National Security Education and Recruitment Strategy for a Long-Term Environment of Limited Resources," *American Intelligence Journal* 31:2 (2013), 23-27; Lowenthal, "Intelligence Education."

[66] Spracher, "National Security Intelligence Professional Education."

[67] Michael Landon-Murray, "Social Science and Intelligence Analysis: The Role of Intelligence Education," *Journal of Applied Security Research* 6:4 (2011), 491-528.

DOI: http://dx.doi.org/10.5038/1944-0472.9.2.1514

organization,[68] and critical and creative thinking.[69] Stephen Coulthart and Matthew Crosston have mapped the constructs of many of America's various degrees in intelligence, grouping curricular content across programs into three categories: core, procedural and domain knowledge.[70]

In addition to conventional approaches to classroom instruction, various kinds of analytic, organizational and operational simulations have become more commonplace, including approaches and programs supported by the ICCAE program.[71] This diversity in approach, perhaps even extending to online instruction, corresponds somewhat to the reality that different people learn differently.[72]

Big Data, Data Science and Analytics Education in the Top Fifty U.S. Universities

Data science, analytics and big data often come together in U.S. academic programs, not surprisingly. Academic degree programs in analytics have been designed chiefly with big data foremost in mind, and in some cases data science and analytics are combined in a single program. For example, at Georgetown University, a Master of Science in Analytics is offered with a concentration in Data Science. The Rensselaer Polytechnic Institute (RPI) also offers a melded data science and analytics program. At Brandeis University, their Master of Strategic Analytics has required coursework in both data science and visualization for big data. About its Analytics program,

[68] Yejun Wu, "Strengthening Intelligence Education with Information-Processing and Knowledge-Organization Competencies," *Journal of Strategic Security* 6:3 (Fall 2013), 10-24.

[69] Michael W. Collier, "Critical Thinking in Academic: What can the U.S. Intelligence Community Expect?" *Journal of Strategic Security* 6:3 Supplement (Fall 2013), 61-64.

[70] Stephen Coulthart and Matthew Crosston, "Terra Incognita: Mapping American Intelligence Education Curriculum," *Journal of Strategic Security* 8:3 (Fall 2015): 46-68.

[71] Allison M. Shelton, "Teaching Analysis: Simulation Strategies in the Intelligence Studies Classroom," *Intelligence and National Security* 29:2 (March 2014): 262-281; Kristan J. Wheaton, Teaching Strategic Intelligence Through Games," *International Journal of Intelligence and CounterIntelligence* 24:2 (March 2011): 367-382; William Costanza, "Building an Intelligence Education Program at Marymount University," *Journal of Strategic Security*, 6:3 Supplement (Fall 2013), 72-79; Melissa Graves, Carl J. Jensen, Walter Flaschka and Carl D. Hill, "Days of Intrigue: Lessons Learned from an Undergraduate Intelligence Case Simulation," *Journal of Intelligence and Analysis* 22:1 (April 2015): 45-60.

[72] Gordon R. Middleton, "A Maturity Model for Intelligence Training and Education," *American Intelligence Journal* 25:2 (Winter 2007/2007), 33-45: James G. Breckenridge, "Designing Effective Teaching and Learning Environments for a New Generation of Analysts," *International Journal of Intelligence and CounterIntelligence* 23:2 (2010): 307-323; David L. Blenkhorn and Craig Si. Fleisher, "Matching Intelligence Teaching Methods with Different Learners' Needs," *Journal of Strategic Security* 6:3 (Fall 2013): 61-72.

the University of Chicago says, "Building from a core in applied statistics, the MScA provides students with advanced analytical training, developing the ability to draw insights from big data."[73] Northwestern's program overview simply starts by saying, "Big Data."[74] The definition provided by Wake Forest actually puts data science synonymous with data analytics: "Data Analytics is a general term for knowledge discovery through the analysis of large datasets, and is also referred to as business intelligence (BI) and data science."[75]

As recently as 2012, there were reportedly few to no academic degree programs in data science, something that has certainly and drastically changed.[76] Taking a look at data science, big data and analytics programs at America's top fifty universities, as ranked by the US News and World Report, it seems intelligence agencies will have no shortage of potential data science hires in the years ahead. And to be sure, these kinds of programs are available at many other universities, and the website Data Science 101 maintains an incredibly comprehensive list of programs in data science, big data and analytics.[77] That list was used as a reference to help ensure that no programs were overlooked, but all programs identified in that list were independently verified by looking at the relevant institutions' own websites. Thus, additional programs not identified by Data Science Community were found and included.

While this study looked only at America's most prestigious national schools, it is important to remember that programs of this type have largely been a graduate phenomenon, and the nation's top technical schools are among the fifty selected for study. This study also excluded those programs with a narrower sector focus, such as business or health analytics, as well as other degrees in computational science and machine learning. Degrees in these areas could, and almost certainly will, provide special value to big data efforts and teams in the IC, and will qualify graduates for IC data science employment. Of course, many graduates of data science and related degree programs may not have deep expertise in relevant substantive areas, and that will need to be addressed. The study also excluded other relevant fields as

[73] The University of Chicago, "Real Data, Real Problems, Real Solutions," available at: *https://grahamschool.uchicago.edu/credit/master-science-analytics/about*.

[74] Northwestern University, "Program Overview: Online Master's in Predictive Analytics," available at: *http://sps.northwestern.edu/program-areas/graduate/predictive-analytics/index.php*.

[75] Wake Forest University, "Certificate in Data Analytics Innovation," available at: *http://charlotte.wfu.edu/certificates-program/certificate-in-data-analytics-innovation-summer-2016/*.

[76] Davenport and Patil, "Data Scientist."

[77] Swanstrom, "Colleges with Data Science Degrees."

DOI: http://dx.doi.org/10.5038/1944-0472.9.2.1514

noted in the data scientist listings at the CIA and NSA, so again, what is yielded is a conservative picture. Nonetheless, thirty-six universities and eighty-three total programs were identified.

Of course, intelligence agencies are not the only organizations and employers who need data scientists, and as noted above, there seems to be a shortage coming. And with probable higher pay in the private sector, it may be tough for intelligence agencies to be competitive. As the CIA's data scientist job description states, however, their data scientists will get to work with information not available elsewhere and in support of U.S. national and homeland security. At the CIA, data scientists earn between $61,444 and $139,523.[78] The listed salary ranges for NSA data scientists begin at $64,923 and go to $103,583.[79] This is compared with an average base pay of $105,395 for data scientists, which suggests the private sector is a more financially alluring destination.[80]

Between doctoral, Master's, Bachelor's and certificate or immersive study, a total of eighty-three programs across fifty universities were identified. It is seen that the emergence of programs in big data, data science and data analytics is occurring primarily at the graduate level, specifically in Master's programs. At that level, eighteen Master's degrees in data science (seven) or analytics (eleven) were identified. Included in the analytics degrees were strategic analytics, predictive analytics, data analytics and government analytics. An additional twenty-five Master's tracks in data science, big data and data analytics were identified. These tracks tend to be found in computer science, engineering, statistics, information science and computational science degrees. At the doctoral level, four relevant programs were identified: a PhD in Big Data at Brown University, a big data concentration in computer science, a PhD in Social Data Analytics at Penn State University, and a track in socio-technical data analytics.

A handful (seven) of schools have introduced undergraduate degrees (ten) in these areas. Dartmouth University, The University of Michigan, University of Rochester, Case Western University and the University of California, Irvine have all established Bachelor of Science degrees in Data Science (the University of Rochester also introduced a Bachelor of Arts in Data Science.) New York University, Penn State and UC Davis offer degrees focused on

[78] Central Intelligence Agency, "Careers & Internships."
[79] National Security Agency, "NSA Careers."
[80] Glassdoor, "25 Highest Paying Jobs in Demand," *Glassdoor Blog*, February 17, 2015, available at: *https://www.glassdoor.com/blog/highest-paying-jobs-demand/*.

analytics. A small number of tracks in data science (three) and analytics (one) can also be found at the undergraduate level.

Lastly, many colleges and universities have established various types of certification or immersive learning programs in data science, analytics and big data (and combinations therein). A total of twenty-two such programs were identified in this study.

Mercyhurst University: Master of Science in Data Science

Mercyhurst has been an early innovator in intelligence education for decades, introducing its first and *the* first American civilian intelligence studies program in 1992. Through the Ridge School of Intelligence Studies and Information Science, the university now offers multiple graduate and undergraduate intelligence degree programs. The newest of these programs is a Master of Science in Data Science. This is the first degree program in the nation to meld data science with intelligence analysis, with one-third of the degree credits coming in the form of intelligence electives. A news article from the University read, "Whereas programs in data science exist, no other American institution offers an equivalent interdisciplinary graduate program that combines both data science and intelligence while also incorporating curriculum from public health and business."[81] The stipulated learning outcomes include:

- Retrieve, organize, combine, and clean data from a variety of private and public data sources
- Store and query data from a variety of private and public data sources
- Apply appropriate techniques to detect patterns and make predictions for private sector and nongovernmental organizations and to support strategic decision-making and action
- Communicate analytic findings in easy-to-understand written, oral, visual, and/or multimedia formats[82]

The core requirements are very much like those required in many data science programs: Machine Learning I and II, Data Science Seminar, Data Visualization, Data Science Tools, Database Technologies, and Research-based Project I and II. This adds up to twenty-four credits, and an additional

[81] Mercyhurst University, "Ridge School Announces Master's Degree in Data Science," December 9, 2014, available at*: http://www.mercyhurst.edu/news/ridge-school-announces-master%E2%80%99s-degree-data-science.*

[82] Mercyhurst University, "Data Science Masters Program," available at: *http://www.theridgeschool.org/ms-data-science.*

twelve credits are completed by selecting among the following sample of courses: Intelligence Theories and Applications, Intelligence and Business Strategy, Geospatial Intelligence, Social Media Analysis, and Public Health Data Analytics, among a number of others. So while there are not formal tracks within the degree, the degree supports any number of career directions for students, from competitive and business intelligence, to national and homeland security, to public health.

The program at Mercyhurst takes two years to complete, so its first graduates will hit "the market" in the spring of 2017. It thus remains to be seen if agencies of the U.S. Intelligence Community think differently about graduates of this program versus the many other data science (and analytics) programs that are now so common in U.S. institutions of higher education. But without question, Mercyhurst has made the determination that, "The data science degree at Mercyhurst was designed to meet a gap in higher education and to enhance the university's competitive edge among liberal arts graduate institutions."[83]

Center for Visualization and Data Science Analytics (CVADA)

Partnerships between government and academia have become a feature of big data and data science activities in universities and colleges, including the Department of Homeland Security's (DHS) Center for Visualization and Data Science Analytics (CVADA). CVADA joins a number of other university-based DHS Centers of Academic Excellence spanning coastal resilience, border, trade and immigration, critical infrastructure resilience, food protection, terrorism, and maritime security, among others. CVADA is a joint effort being implemented by Rutgers University and Purdue University. Broadly, CVADA's mission is to help create "the scientific basis and enduring technologies needed to analyze large quantities of information to detect security threats to the nation."[84] At CVADA,

> "Researchers and educators develop faster ways for data to be collected, distilled, managed, visualized, understood, and shared before, during, and after a crisis. CVADA is creating a foundation in visual and data analytics to enable swiftly sifting through a tsunami of

[83] Mercyhurst University, "Ridge School Announces Master's Degree in Data Science."
[84] U.S. Department of Homeland Security, "Welcome to the Centers of Excellence," available at: *http://www.dhs.gov/science-and-technology/centers-excellence*.

> information, in diverse forms, to get early warning of potential threats."[85]

CVADA's broad areas of research include:

- Public Safety Coalition Projects
- Enterprise Resiliency Experiments
- Sports Evacuation Planning
- Visual Analytics for Security Applications
- Information and Gathering Distillation
- Information Networks and Analysis
- Information-Driven Modeling and Simulation
- Information-Driven Decision Making
- Education Initiatives
- International Collaborations[86]

Each university partner operates different facets of CVADA. Rutgers University houses the Command, Control, and Interoperability Center for Advanced Data Analysis (CCICADA). Purdue University runs the Visual Analytics for Command, Control, and Interoperability Environments (VACCINE). CCICADA's projects and work include:

- Border Security (Port Security, Unaccompanied Alien Children)
- Cyber Attacks
- Disease Surveillance
- Flood Mitigation (Raritan River, Hurricane Sandy)
- Prevention of Child Sex Trafficking
- Nuclear Detection
- Stadium Security
- Urban Security
- U.S. Coast Guard (Arctic Strategy, Fisheries Enforcement, Resource Allocation)[87]

In the realm of cyber security, CCICADA recently delivered a "Report on the State of Cyber Security Education in the US" to the Cyber Security Division of DHS. This is one of many such reports and papers that have been produced

[85] Homeland Security University Programs, "Center for Visualization and Data Analytics (CVADA)," available at*: https://cvada.hsuniversityprograms.org/centers-of-excellence/cvada/.*

[86] Ibid.

[87] Command, Control and Interoperability Center for Advanced Data Analytics, "Research," available at: *http://www.ccicada.org/research/*.

under the auspices of CCICADA. In addition to workshops and other events, there is also an educational component of CCICADA, including a summer research program for undergraduates and a graduate fellowship that funds students for two to three year periods. Students who receive these fellowships study topics related to "homeland security decision-making and data analysis."[88] CCICADA's work is done in conjunction with seventeen established partner institutions, a majority of them being other universities.

Purdue's VACCINE "designs tools for the seven DHS components, which include Federal Emergency Management Agency, U.S. Citizenship & Immigration Services, U.S. Coast Guard, U.S. Customs & Border Protection, U.S. Immigration & Customs Enforcement, U.S. Secret Service, and Transportation Security Administration."[89] The applications and technologies developed at VACCINE are meant to help stakeholders effectively and efficiently respond to natural and manmade crises, whether a terrorist attack or pandemic, by organizing—analytically and visually—immense data flows in an actionable, accessible way.

Like CCICADA, VACCINE has a large network of partners which spans academia, law enforcement, the private sector, and other parts of the U.S. Intelligence Community. The researchers at VACCINE develop "innovative interactive visualization, analysis, and decision-making tools,"[90] and the work and tools of VACCINE generally fall into the following categories:

- Investigative Analysis and Anomaly Detection
- Trend Identification and Predictive Analytics
- Spatiotemporal Exploration and Visual Analytics
- Risk-Based Decision Making and Resource Allocation
- Image/Video Analytics and Recognition

Like CCICADA, VACCINE has an educational dimension to it as well. Their educational mission statement is as follows:

> "VACCINE's mission is to educate current homeland security stakeholders and the next generation of talent in effective development

[88] Command, Control and Interoperability Center for Advanced Data Analytics, "CCICADA Fellowship Program," *available at: http://www.ccicada.org/2014/05/13/ccicada-fellowship-program/*.

[89] Visual Analytics for Command, Control, and Interoperability Environments, "Research," available at: *https://www.purdue.edu/discoverypark/vaccine/research/index.php*.

[90] Ibid.

> and use of visual analytics systems. Our educational initiatives span the career development pipeline ranging from undergraduate and graduate level work to professional education and training programs. Our goal is to build a diverse, highly capable, technical workforce for the Department of Homeland Security enterprise by administering various programs and initiatives at our center, partner research institutions, and minority-serving institutions."[91]

At the undergraduate level, this mission manifests in summer research programs. At the graduate level, fellowships and scholarships are available to students, and visual analytics courses are available at a number of VACCINE partner universities. A number of professional development opportunities are also available in the form of webinars and faculty workshops.

Discussion

Like any number of other sectors, the U.S. Intelligence Community must exercise due diligence in identifying the opportunities, uses, limits and issues of the era of big data. Academics, both in data disciplines and intelligence studies, will also have a continuing role in helping answer a range of critical questions. What are the new insights and analytic products that are made possible? How can existing analysis and products be bolstered by the inclusion of data science support? What cultural and workplace dynamics—and other organizational constraints—in the IC might undermine the efficacy of big data and the work of data scientists, and how can those issues be managed to a positive end? Grappling successfully with such issues will help the IC attract and retain the best—or at least better—data scientists and ensure that it is maximizing big data. Conversely, what intelligence topics and questions are unlikely to be helped by big data analytics? How can data scientists still be potentially utilized in such circumstances? The ethical and legal dimensions must also be given a central place among practitioner, scholarly and public dialogue.

Answering many of the above questions—with the support of a competent corps of data scientists—will also help the IC avoid big data investments and initiatives that may prove to have little benefit. Whatever information asymmetries exist between the IC and outside vendors and industry need to

[91] Visual Analytics for Command, Control, and Interoperability Environments, "Education," available at: *http://www.purdue.edu/discoverypark/vaccine/education/index.php*.

be addressed so that intelligence agencies can be smart consumers of data science and all the products floating around big data.

This article begin to explore these questions, in the process laying out what the CIA and NSA look for in big data and data scientists. Human capital dimensions of data science are of course being worked out in the IC, as elsewhere, just as the technology and insights themselves are evolving. The ability of the IC to attract and hire individuals with data science and other related backgrounds will be key to the optimization and management of big data intelligence. Data scientists who enter the IC will have the task of exploring all the potential uses and benefits of this explosion in data and technology. Imaginative and creative minds will thus be just as important—and potentially more important—than in conventional intelligence analysis.

The IC is looking to many disciplines to draw its data scientists from, and in the coming years, there will be many more potential hires with data science and closely related degrees. In the job descriptions provided by the NSA and CIA, both the art and science sides of data science are represented. These agencies are looking for creative individuals who are also technologically savvy so that they may build sound and inventive algorithms and models. The NSA considers its data scientists part computer scientists part analysts, and expects its data scientists to "Develop analytic plans, engineer supporting algorithms, design and implement solutions which execute analytic plans."[92] This is an important dichotomy that must be reflected in data science education, which will probably come down to a class-to-class, instructor-to-instructor basis. Mechanics are surely easier to teach than creativity, but an effective IC data scientist will very much need both facets. As Rachel Schutt has commented, today's data science students must be trained for tomorrow's problems,[93] which may have little connection with today's data and tools.

In addition to realizing the optimal use of big data, IC data scientists will also exercise a degree of responsibility to act as ethical stakeholders. As Omand et al. point out in the context of social media intelligence, it is incumbent on intelligence and policy officials that the uses of big data are both necessary and legitimate.[94] In addition to privacy violations, misuse and subsequent public and political backlash may result in the placement of limits on big data

[92] National Security Agency, "NSA Careers."
[93] Schutt, "Educating the Next Generation of Data Scientists."
[94] David Omand, Jamie Bartlett and Carl Miller, "Introducing Social Media Intelligence (SOCMINT)," *Intelligence and National Security*, 27:6 (December 2012): 801-823.

that exceed what is necessary, limiting its value to intelligence and security missions beyond what is required.

Many of the skills and attributes—curiosity, creativity, energy, and the ability to grow and adapt—that make for a robust data scientist seem highly relevant to the conventional work of intelligence analysts. It should be investigated whether introductory courses in big data and data science can activate or accentuate skills and abilities that can help other intelligence analysts do a better job. Writing way back in 1957, Washington Platt observed that perhaps more important than statistical formulas, intelligence officers should be trained in statistical thinking. Today, that partly includes understanding big data and data science, which despite some rather grandiose claims, are built on statistical foundations and will be subject to longstanding data problems and limits.[95] As Tim Harford has written, "The promise that 'N=All,' and therefore that sampling bias does not matter, is simply not true in most cases that count."[96] Finding biases in huge and messy data sets will be no easy feat.[97]

Instruction in big data analytics will also put intelligence analysts in a better position to engage and work with their data science colleagues in the IC. Building these foundations can enable analysts to grasp what they are being told from IC data scientists, but also to ask probing questions of those data scientists and help communicate big data insights to policymakers. In this vein, James Madison University's Bachelor of Science in Intelligence Analysis includes courses in data science, mining and visualization. However, it seems these remain rare offerings in academic intelligence curricula. Conversely, data scientists also need relevant domain expertise. This suggests that the graduates of Mercyhurst University's MS in Data Science will be in an advantageous position when compared to their counterparts from more general data science and analytics programs. While not a standalone degree like Mercyhurst's, Johns Hopkins University offers a Government Analytics Master's program that can be paired with a graduate certificate in intelligence.

While America's top universities are responding to this demand with remarkable speed and quantity, it remains to be seen if the public and private sectors are able to fill their data scientist openings to their satisfaction. If

[95] Platt, Washington, *Strategic Intelligence Production: Basic Principles* (New York: Frederick A. Praeger, 1957).
[96] Harford, "Big Data."
[97] Ibid.

https://digitalcommons.usf.edu/jss/vol9/iss2/6
DOI: http://dx.doi.org/10.5038/1944-0472.9.2.1514

projections on demand are accurate, the rapid growth in programs may not be keeping pace with job demand. Further, the IC may experience difficulty in attracting data scientists, who can start in the private sector with probably higher wages and no year-long security background check. The information and insight needed to gauge whether the IC is meeting its human capital needs in the area of data science, qualitatively and quantitatively, will probably remain unavailable to researchers in a direct fashion.

In addition to academics, universities are making contributions, if indirect, to the security and intelligence communities through the research they are conducting. A majority of the universities surveyed for this study had some sort of formal activity around big data and data science, and like the security-focused big data research centers discussed above, some have relationships with the federal government (such as the Regional Big Data Innovation Hubs). And while the research in these entities may not directly inform the work of intelligence agencies, there is surely an indirect or trickledown benefit. The sheer number of academic journals in big data and data science also indicates that much academic work is being done.

University centers with a security or intelligence focus can help the IC overcome a number of issues, creating mechanisms that bring to bear external advances while “reading in” outside specialists who can then work toward tools that serve unique intelligence and security functions. Relying on organizations like IARPA is probably not enough.

The IC Centers for Academic Excellence grant program may represent another opportunity for the IC to support big data research and education with an intelligence bent, helping to establish research programs and additional academic programs like the one found at Mercyhurst University. As the program stands, it is largely responsive to what universities propose and does not highlight big data or data science as an existing need or focus.

Conclusion

This article has been an effort to explore data science applications, issues and human capital dimensions in the U.S. Intelligence Community. Much, much more needs to be done. Big data and data science are evolving areas, and new uses and technologies will be produced with great regularity in the coming years. Human capital will be the key to the IC best utilizing those uses and technologies, and successfully filling these ranks is perhaps the most immediate and important challenging facing the IC. This article thus examined more closely data science human capital dynamics in the IC, looking first at the issue areas and "accounts" that big data does and is likely to speak to, and then breaking down data science job listings in the IC. After all, perhaps the best, simplest way to approximate 'data science' is probably to look at "what data scientists do." To fill the ranks of data scientist, the IC will be looking to higher education, which has only recently begun introducing data science programs. However, as this article found, higher education has responded very quickly, with over half of America's top national universities offering programs in data science, big data and/or data analytics. Thus far, only one program has emerged that focuses on the confluence of data science and intelligence analysis, and data science has not really found its way into academic programs in intelligence studies and intelligence analysis. University-based research centers can also help the IC illuminate new technologies and applications in this realm, as demonstrated by those at Purdue and Rutgers. It remains to be seen if the IC feels its human capital demands are being adequately supplied, quantitatively and qualitatively, but this article has done its best at showing what human capital needs are in the IC, and how U.S. higher education is responding in the way of new academic and research programs.