Article accepted for publication in *International Affairs* (2017)

Beyond the Buzzword: Big Data and National Security Decision-making

DAMIEN VAN PUYVELDE, STEPHEN COULTHART AND MAHMUD SHAHRIAR HOSSAIN

This article explores the role big data play in the national security decision-making process. The global surveillance disclosures initiated by former NSA contractor Edward Snowden have increased public and academic discussions about big data and national security. Yet, efforts to summarize and import insights from the vast and interdisciplinary literature on data analytics have remained rare in the field of security studies. To fill this gap, we explain the core characteristics of big data, provide an overview of the techniques and methods of data analytics, and explore how big data can support the core national security process of intelligence. We find that data analytics tools contribute to and influence all the core intelligence functions in the contemporary US national security apparatus. However, these tools cannot replace the central role of humans and their ability to contextualize security threats.

In the last decade, big data has been an ubiquitous buzzword in academic and professional circles and in the media. Some commentators have praised big data as 'the new oil of the 21st century', 'the world's most valuable resource' and 'the foundation of all of the megatrends that are happening today, from social to mobile to the cloud to gaming'.[1] The growth of big data analytics can be explained from a market-based perspective. On the supply side, data have become more readily available and processing power has kept increasing – as predicted by Moore's Law in the 1970s. Rapid advances in instrumentation and sensors, digital storage and computing, communications and networks, including the advent of the Internet in the 1990s, have spurred an ineluctable march towards the 'big data revolution',[2] generating and giving access to more and more data. Every day, humans directly or indirectly create 2.5 trillion megabytes of data.[3] As increasingly large amounts of data are captured from humans, machines,

[1] Peter Sondergaard, 'Big Data Fades the Algorithm Economy', *Forbes*, 14 Aug. 2015; Chris Lynch, cited in Cisco, '5 reasons why your data center is everywhere', http://www.cisco.com/web/global/assets/pdf/dc-05-data-center-is-everywhere-top5-infograph-cte_enuk.pdf; Editor, *The Economist*, 6 May 2017, cover page.
[2] See: Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data. A Revolution That Will Transform How We Live, Work and Think*, London: John Murray, 2013; Xindong Wu et al., 'Data mining with big data', *IEEE Transactions on Knowledge and Data Engineering* 26: 1, 2013, p. 97.
[3] IBM, 'Bringing Big Data to the Enterprise', https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

and the environment, the temptation to analyse them grows, a phenomenon sometimes known as datafication.[4] The current deluge of data, spurred by the increased digitization of information,[5] provides countless opportunities for data mining, a set of techniques seeking to extract hidden patterns from datasets, in a variety of contexts.[6] These new capabilities have started affecting organizations and the core processes they follow.

The big data craze has also gained traction within government, including in intelligence agencies, which have always relied on data sources to collect raw information and develop products for the consumption of a host of decision-makers. In the US intelligence community, big data has become institutionalized through the establishment of advanced analytics units in civilian and military intelligence agencies, and a growing number of data analytics projects funded through organizations like the Intelligence Advanced Research Project Activities (IARPA), the Defense Advanced Research Project Agency (DARPA), or the Central Intelligence Agency's technology incubator In-Q-Tel.[7]

Interest in data analytics has been growing due to the demand for more valid intelligence products following the controversies caused by the 9/11 attacks and the absence of weapons of mass destruction (WMD) in Iraq. Prior to 9/11 the US intelligence community lacked and missed specific pieces of information pointing to the terrorist plot. In 2002, a national intelligence

[4] Kevjn Lim, 'Big Data and Strategic Intelligence', *Intelligence and National Security* 31: 4, 2016, p. 622.
[5] Digitization converts analog content into digital information. Honovar notes that 'anything that is describable can be described using a computer program'. See: Vasant G. Honavar, 'The Promise and Potential of Big Data: A Case for Discovery Informatics', *Review of Policy Research* 31:4, 2014, p. 327.
[6] Gordon E. Moore, 'Cramming More Components onto Integrated Circuits', *Electronics*, 1965, pp. 114–117; David R. S. Cumming, Stephen B. Furber, Douglas J. Paul, 'Beyond Moore's law', *Philosophical Transactions of the Royal Society* 372, 2014, pp. 1–2; Frans Coenen, 'Data Mining: Past, Present, and Future', *The Knowledge Engineering Review* 26: 1, 2011, p. 25.
[7] John Poindexter, 'DARPA's Initiative on Asymmetric Threat: Total Information Awareness', *DARPATech 2002 Symposium*, https://w2.eff.org/Privacy/TIA/darpatech2002/slides/PoindexterIAO.pdf; Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity, 'Incisive Analysis', https://www.iarpa.gov/index.php/about-iarpa/incisive-analysis, and 'Anticipating Surprise', https://www.iarpa.gov/index.php/about-iarpa/anticipating-surprise; Defense Advanced Projects Research Agency, 'XDATA', http://www.darpa.mil/program/xdata; In-Q-Tel, 'Portfolio', at https://www.iqt.org/portfolio/.

estimate made a series of erroneous assessments regarding Iraq's WMD program, which were later used to justify the US decision to go to war in Iraq.[8] These events casted doubt on the intelligence collection and analysis capabilities of the US government, especially in the domain of human intelligence (HUMINT), and increased the pressure on senior decision-makers to adapt intelligence processes to an increasingly complex security environment. Big data capabilities, it was hoped, would compensate the limitations, and sometimes the absence, of HUMINT. Consequently, US intelligence agencies began to embrace more systematic and sophisticated data collection and analysis techniques.

Given the widespread use of the term big data, one would expect to find a sophisticated account of what it means, what it does, and how it works in the national security context. However, the field of security studies has, thus far, paid little attention to this concept.[9] Scholars have tended to focus on issues of privacy and liberties, following the revelations made by former NSA contractor Edward Snowden about bulk data collection programs deployed by the US and its five eyes partners.[10] Many existing accounts provide brief overviews of contemporary technological capacities to collect and compute vast amounts of data, but few delve into what is meant by big data in a variety of security contexts.[11] The absence of a comprehensive study on

[8] Thomas Kean, *The 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks upon the United States* 2004, p. 361; Laurence H. Silberman and Charles S. Robb, *The Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction*, 2015, p. 2; United States Congress, Intelligence Reform and Terrorism Prevention Act of 2004, Public law 108-458, 17 Dec. 2004.

[9] For some attempts to consider the role of data analytics in national security, see: David Omand, Jamie Bartlett and Carl Miller, Introduction Social Media Intelligence (SOCMINT), *Intelligence and National Security* 27: 6, 2012, pp. 801–823; Lyria Bennett Moses and Janet Chan, 'Using Big Data for Legal and Law Enforcement Decisions: Testing the New Tools', *University of New South Wales Law Journal* 37, 2014, pp. 643–678; Lorna Mui, 'Transparent Fictions: Big Data, Information and The Changing Mise-en-Scene of (Government and) Surveillance', *Surveillance and Society* 13: 3-4, 2015, pp. 354–369. The trend is also visible on the program of the International Studies Association Annual Convention, which featured panels on 'National Security and Intelligence in a Big Data Age: Innovation, Ethics And Legitimacy' and 'Bid Data and National Security' in 2015 and 2016.

[10] See for example: Louise Amoore, 'Security and the Claim to Privacy', *International Political Sociology* 8, 2014, pp. 108–12.

[11] See for example: Anders Koed Madsen et al., 'Big Data: Issues for an International Political Sociology of Data Practices', *International Political Sociology,* 2016, pp. 1–22; Chris Pouling, 'Big Data Custodianship in a Global

big data and national security decision-making is problematic because it limits researchers' ability to consider the implications of the big data 'revolution' in the field of security. In a recent article on the role of the Internet in violent extremism and terrorism, Maura Conway openly recognizes that she hesitates to use the term big data, possibly due to its conceptual ambiguity.[12] This example is symptomatic of the broader need for conceptual clarity on big data in the fields of security and international affairs.

This article explores and clarifies what big data means and what roles big data tools play in national security decision-making. Two main objectives motivate this article. First, we want to integrate multi-disciplinary research on big data more comprehensively into the social scientific study of security to develop a common understanding of its role and limits. Without such conceptual clarity, research in the field is likely to follow inconsistent and disjointed paths. Second, developing a common understanding of the role and limits of data analytics will facilitate its effective use by security practitioners and decision-makers. The latter will, understandably, be reluctant to accept the results of automated analysis of big data if they do not understand the process behind the key findings they are briefed, and cannot convincingly explain their resulting decisions to the public. Conversely, decision-makers may put undue confidence in big data tools, mistakenly construing technological solutions as a silver bullet that can help them overcome the complex dilemmas they face.

Society', *SAIS Review* 34: 1, 2014, pp. 109–116; Charles J. Dunlap Jr., 'The Hyper-Personalization of War: Cyber, Big Data, and the Changing Face of Conflict', *Georgetown Journal of International Affairs*, 2014, pp. 108–118; Kenneth Cukier and Viktor Mayer-Schoenberger, 'The Rise of Big Data: How It's Changing the Way We Think About the World', *Foreign Affairs* 92: 3, 2013, pp. 28–40. For a notable exception, see the recent special issue on *Securing with algorithms: Knowledge, Decision, Sovereignty* edited by Louise Amoore and Rita Raley in *Security Dialogue* 48: 1, 2017.

[12] Mayra Conway, 'Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research', *Studies in Conflict & Terrorism* 40: 1, 2017, pp. 77–98.

The article begins by considering what constitutes big data. While explicit definitions of big data are rare in security studies, other fields like computer sciences, computer engineering, information sciences and business administration have already produced a vast literature on the subject.[13] To take stock of the diverse perspectives on the subject, we rely on De Mauro and her colleagues' effort to define big data thanks to a survey of 1,581 conference papers and journal articles on the topic. The resulting definition considers big data as: 'the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value'.[14] We describe and explain how the components in this definition – the characteristics of data (high volume, velocity, variety, and also veracity), technology and analytical methods – can be understood in the context of national security. The final section explores the value of big data in the national security context, through the prism of core intelligence functions. We explore how big data affect intelligence requirements, collection, processing and exploitation, analysis, dissemination, and counterintelligence and security. We conclude that, while the growth of big data analytics is changing the character of national security processes, the human nature of these processes remains unchanged. Given the growing volume and velocity of data inflow in the national security process, some degree of machine learning and artificial intelligence decision-making is inevitable to help prioritize analysis. However, automated analysis of data is not and will not fundamentally alter the need for human judgment at multiple levels in national security decision-

[13] See for example: Andrea De Mauro, Marco Greco, and Michele Grimaldi, 'What is big data? A consensual definition and a review of key research topics', in Georgios Giannakopoulos, Damianos P. Sakas, and Daphne Kyriaki-Manessi , eds., *AIP conference proceedings* 1644: 1, 2015, pp. 97–104; Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips, '"Big Data": big gaps of knowledge in the field of internet science', *International Journal of Internet Science* 7: 1, 2012, pp. 1–5.

[14] Andrea De Mauro, Marco Greco, and Michele Grimaldi, 'What is Big Data? A Consensual Definition and a Review of Key Research Topics', International Conference on Integrated Information, AIP Conference Proceedings, 2015, p. 103.

making. Consequently, social scientific research on big data should focus on adapting human-machines interactions to make them as effective as possible.

The Characteristics of Big Data: Volume, Velocity, Variety, and Veracity

The expression big data is often understood as a set of very large datasets. But what exactly qualifies as a very large dataset? Volume is understood and processed differently in multiple fields and at different points in time. For a social scientist, a dataset including hundreds of thousands of entries may seem large, but not so much for a computer scientist. Similarly, while computer scientists might have considered a database of hundreds of thousands of entries to be very large in the early days of computing, today's researchers work with billions of entries. A 2010 study found that the amount of data produced globally was 1.2 zettabytes, or 1,200,000,000,000 trillion gigabytes. By the year 2020 it is expected that worldwide data production will reach 35 zettabytes.[15] The desire and ability to process such large volumes of data is a significant component of the definition of big data, but volume alone is not sufficient to define big data.

Early definitions of big data describe how large amounts of data put heavy demands on computing power and resources thus causing a 'big data problem'.[16] As the world keeps producing more and more data, this problem has far from disappeared. Processing capabilities for all these data now lag behind storage capabilities. In other words we have access to massive amounts of data but are not able to use all of them.[17] Volume, therefore, should not be considered

[15] John Gantz and David Reinsel, 'The Digital Universe Decade – Are You Ready?', IDC, May 2010, p. 1, https://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf.
[16] Michael Cox and David Ellsworth, 'Application-controlled demand paging for out-of-core visualization', Proceedings of the 8th conference on Visualization, 1997, p. 235.
[17] Neil Couch and Bill Robins, *Big Data for Defence and Security* (London: Royal United Services Institute: 2013), p. 32.

on its own, but in relation to the ability to store and process data. Big data are not only understood as numbers or volume, but are defined in relation to the capacity to use these data. Following this approach a common definition of big data describes 'datasets whose size is beyond the analytical capacities of most database software tools'.[18] One national security professional explains that 'big data' starts when Excel is not enough anymore.[19]

This capacity to use data is not only compounded by increasingly large volumes of data, but also by data velocity. The speed at which new data are generated and change is increasing, which poses further storing and processing challenges. Twitter users, for instance, generate on average 6,000 tweets per second.[20] To cope with the velocity of data, researchers and intelligence practitioners have sought to combine multiple data streams. In one such project, the National Security Agency combined data from 'phone conversations, military events, road-traffic patterns, public opinion—even the price of potatoes'.[21] In this case, automated data analysis partially replaces humans, who could not process all these data in a timely fashion. According to a former official with knowledge of the program, analysts found that introducing more data into the program also led to more accurate predictions of where insurgent attacks would occur.

The 'big data problem' is further complicated by the growing variety of data. In computer sciences, data are generally considered as alphanumeric characters and symbols that are stored, processed and transmitted. In the last decades, the variety of data available to researchers exploded and technology provided the means to tap into new data sources. The proliferation of smartphones and wearable technology connected to the Internet and equipped with audio and

[18] Connie L. McNeely and Jong-on Hahm, 'The Big (Data) Bang: Policy, Prospects, and Challenges', *Review of Policy Research* 31: 4, 2014, p. 305.
[19] National security professional, conversation with authors, April 2016.
[20] Internet Live Stats, 'Twitter Usage Statistics', at http://www.internetlivestats.com/twitter-statistics/.
[21] Siobhan Gorman, Adam Entous, and Andrew Dowell, 'Technology Emboldened the NSA', *Wall Street Journal,* 9 June 2013, at http://www.wsj.com/articles/SB10001424127887323495604578535290627442964.

video sensors as well as GPS locators is only one example of how technological advances produce more and more diverse data. Humans and their environment have always produced data, for example biological and meteorological data, but a growing number of sensors now collect and store these data in increasingly greater quantity and quality. In addition, the advent of cyberspace has led to new types of data, generated by networks and the humans who use these networks. Surfing on the World Wide Web generates digital trails of data that can be accessed by social media sites like Twitter and search engines like Google. The computer networks that constitute the digital layer of cyberspace also generate data like web server logs.[22] Each of these data types, some older than others, poses specific challenges related to their volume and velocity, thus contributing to the increasing variety of data available for analysis.

Researchers usually classify data in three categories: structured, semi-structured, and unstructured. Structured data, writes one observer, have been reformatted and "organized into a data structure so that elements can be addressed, organized and accessed in various combinations to make better use of the information'.[23] In other words, structured data have been processed so that they are easily stored for retrieval and analysis. Examples of structured data include texts and numeric information that are stored in traditional relational databases, meaning the data can fit in rows and columns. Quantitative social scientists often use structured data in the form of Excel spreadsheets. Structured data are the least common type of data but the most commonly analysed.

Semi-structured is the next most common data type and is harder to store and analyse than structured data. This type of data is structured because a certain number of attributes

[22] Aaron F. Brantly, 'Changing the Game: Cyberspace and Big Data Driven National Security Intelligence', in Damien Van Puyvelde and Aaron F. Brantly, *US National Cybersecurity: International Politics, Concepts and Organization* (London: Routledge, 2017), p. 141.
[23] Margaret Rouse, 'Semi-structured data', http://whatis.techtarget.com/definition/semi-structured-data.

represent each data-object. For example, a Microsoft Word document with headed sections conveys semi-structured data. While this document lacks the rigid organization of structured data, it contains a number of attributes including chapters or section headings, tags informing users about the date of creation, and revisions of the document, which can help users organize and analyse it.[24] Unstructured data is the fastest-growing category and the most common. It comes in a variety of formats including audio, video, analog data, books, images, web pages and more. The proliferation of social media platforms and personal devices has increased the amount of publicly available unstructured data. This type of data poses a significant challenge to data scientists inside and outside of modern intelligence agencies. The former chief technology officer for the Central Intelligence Agency, sums up the problem for the US intelligence community: 'Our data is always fragmented, and we're trying to make sense of fragmented data options, which is extremely difficult…how we analyze every piece of data, how we reprocess it to continue to make better sense of what is going on – that is the biggest [challenge] we have, especially when we can't get complete databases'.[25] Large organizations, such as Google and multiple government intelligence agencies increasingly rely on novel types of databases that can store a wide variety of data types from structured to unstructured, and organize them for subsequent analysis.

Veracity is another characteristic that is relevant to consider as we discuss the use of big data in national security. This characteristic refers to 'the biases, noise and abnormality in data'.[26] In the national security context, more than in any other fields, data and information should be approached sceptically because adversaries often actively alter data to deceive and mislead.

[24] Ibid.
[25] Frank Konkel, 'The intelligence community's big-data problem', FCW, 1 March 2014, https://fcw.com/articles/2014/03/13/ic-big-data.aspx.
[26] Kevin Normandeau, 'Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity', 12 Sept. 2013, http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.

National security professionals must question the representativeness and validity of data collected on social media, for instance.[27] Questioning the veracity of data encourages the consumers of big data to consider whether the data that are being stored and mined are reliable and meaningful enough to answer the problems at hand.[28] A classic example of how data veracity can affect national security decisions is the case of Robert McNamara's tenure at the Pentagon during the Vietnam War. While not known as big data at the time, McNamara oversaw a large-scale effort to collect data on the US war effort in Vietnam. His team of 'Whiz Kids' applied statistical techniques honed in business to understand the war and assess the effectiveness of US decisions. Ultimately, however, McNamara put undue confidence on key metrics that he used to assess the success of US efforts in Vietnam, in particular body count data on the enemy. According to one General, these data were susceptible to being distorted by commanders who wanted to demonstrate their effectiveness.[29] McNamara himself notes in his memoirs that data used by the US military was 'inflated by the considerable falsification of data submitted by South Vietnamese officials'.[30] A more recent example of intentional data manipulation is provided by Gary King and his colleagues, who estimate that the Chinese government fabricates 448 million social media posts a year.[31] Given its volume and diversity, big data are inherently noisy, which reinforces the difficulty in identifying signals, or genuine posts in this case, from

[27] See Zeynep Tufecki, 'Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls', Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014.
[28] Normandeau, 'Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity'.
[29] Kenneth Cukier and Viktor Mayer-Schönberger, 'The Dictatorship of Data', *MIT Technology Review*, 31 May 2013.
[30] Robert S. McNamara with Briand VanDeMark, *In Retrospect. The Traget and Lessons of Vietnam* (New York: Vintage Books, 1996), p. 104.
[31] Gary King, Jennifer Pan, and Margaret Roberts, 'How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument', Forthcoming Working Paper, 7 April 2017.

the background noise.[32] The risk, if the veracity of big datasets cannot be established, is that they will mislead analysts, and possibly decision-makers.

The Technology and Analytical Methods of Big Data

Big data are not only about data, but also about the means, both technological and methodological, by which data are stored, processed and analysed. A complete review of the technology driving big data is beyond the scope and purpose of this article. Instead we focus on select capabilities to illustrate the technology and methods of big data.

Processing vast amounts of data requires robust hardware and software capabilities. Hadoop, to focus on one example, is 'an open source framework that enables the distributed processing of big quantities of data by using a group of dispersed machines and specific computer programming models'.[33] This framework relies on parallel processing to help users process and analyse very large datasets by leveraging the computing power of hundreds or even thousands of computers simultaneously. Hadoop is revolutionary because it facilitates the computation of massive amounts of data without relying on expensive super computers. Its ability to distribute processing makes big data capabilities more accessible to institutions and businesses that have limited resources. This capability did not go unnoticed in the intelligence community, and in 2009, the NSA announced that it was using a 'new system by linking its various databases and using Hadoop software' to amplify computing power and analyse data.[34] Smaller intelligence agencies, like the French *Direction Générale de la Sécurité Extérieure* and

[32] Roberta Wohlstetter, *Pearl Harbor: Warning and Decision* (Stanford: Stanford University Press, 1962), p. 3.
[33] De Mauro, Greco, and Grimaldi, 'What is Big Data?', p. 99.
[34] Gorman, Entous, and Dowell, 'Technology Emboldened the NSA'.

the British Government Communications Headquarters use similar software to improve their ability to access and process big data.[35]

The application of data processing and storage technology would be useless without a way to extract information from data. Data are nothing more than a series of symbols, and only become information when they are processed to generate meaning.[36] Information, in turn, helps analysts answer who, what, where, when, and how questions to support decision-makers.[37] Data analytics relies on algorithms, defined as sets of rules or actions to be performed to process data. Big data processing capabilities are generally divided into two main types: data management and analytics. Data management includes the processes and technologies seeking to acquire and record data (intelligence collection), to clean, annotate and represent data so that it is ready for analysis (processing). Analytics techniques help process data to extract information from them (exploitation).[38] Analytics can be applied to multiple data sources including texts, audio and video records, for instance. Here users hope to learn from a very large body of data, phenomena that they could not identify or comprehend using only smaller amounts.[39] To process the data and support data management and analytics, algorithms mine vast troves of data from which they extract information for human consumption, an activity known as data mining. Algorithms are also used to look for patterns that will be used by computers to adjust specific program actions.

---

[35] Editor, 'The DGSE', 2 May 2014, *Intelligence Online*, https://www.intelligenceonline.com/government-intelligence_organizations/2014/02/05/the-dgse,108006709-BRE; Government Communications Headquarters, 'Applied Research. Advancing the Art of the Possible', https://www.gchqcareers.co.uk/departments/applied-research.html.

[36] Stephen Gary and Randy Borum, 'Evolving Cyber Intelligence', in Damien Van Puyvelde and Aaron F. Brantly, eds, *US National Cybersecurity: International Politics, Concepts and Organization* (London: Routledge, 2017), p. 123.

[37] Note that the 'why' question is not mentioned. This is because big data analytics can reveal correlation, not causation. On the limits of big data in this context, see: William Roberts Clarm and Matt Golder, 'Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?' *PS: Political Science & Politics* 48: 1, 2015, p. 66.

[38] Amir Gandomi and Murtza Haider, 'Beyond the hype: Big data concepts, methods, and analytics', *International Journal of Information Management* 35, 2015, p. 140.

[39] Cukier and Mayer-Schoenberger, 'The Rise of Big Data', p. 28.

This process is called machine learning because the computer learns from a set of data labelled as examples in a process called 'training', and adjusts its algorithm to assign values or categories to unlabelled examples.[40]

Data scientist Brandon Rohrer groups machine learning algorithms in five families based on the type of question they answer. First, two-class or binary classification algorithms help answer questions that only have two possible answers like yes or no, on or off, and eventually suspicious or normal. One such question could ask, is this a picture of Osama bin Laden or not? More advanced algorithms perform multi-class classification, answering questions that have several possible answers. Such questions might include: what is the topic of this article? What is the mood of this social media post? Who is the speaker in this recording? A second family of algorithms performs anomaly detection and identifies data points that are not normal, or outliers. This can be used to answer a variety of questions including 'Is this Internet message typical? Is this combination of purchases very different from what this customer has made in the past?' Third, regression algorithms can help answering numeric questions starting with 'how much' and 'how many'. This capability can be used to make predictions, for example asking how many followers a specific social media profile will get in the next week or month. Making such a prediction can, for instance, help national security professionals prioritize counter-propaganda efforts. Regression algorithms can also conduct classification to rank or compare events, objects, preferences and persons, and answer questions of likelihood. A two-class classification (logistic) regression could help answer: 'how likely is this employee to be an insider security threat?'

[40] Margaret Rouse, 'Machine Learning', *TechTarget*, at http://whatis.techtarget.com/definition/machine-learning. For example see: Hao Wu, Michael Mampaey, Nikolaj Tatti, Jilles Vreeken, M. Shahriar Hossain, Naren Ramakrishnan, 'Where Do I Start? Algorithmic Strategies to Guide Intelligence Analysts', Proceedings of the ACM SIGKDD ISI-KDD, 2012), Article No. 3.

Fourth, unsupervised learning seeks to answer questions about how data are organized. Here machine learning can perform clustering, separating data into intuitive groups. Big data tools can, for instance, organize documents or social media posts into topic groups, or identify which groups of social media users like the same kind of extremist propaganda. Another type of unsupervised learning summarizes, simplifies, or condenses large datasets to understand what groups of factors vary together. For example, big data tools can automatically construct a textual summary of hundreds of thousand documents containing descriptions of extremist propaganda, or generate a summary of each topic group of social media posts everyday, instead of employing human analysis to sift through millions of Facebook posts and tweets to detect suspicious activities. Finally, reinforcement machine-learning algorithms process large amounts of data and choose actions, among a preselected pool of possibilities.[41] These algorithms gather data from the environment and learn an ideal behaviour based on trial and error. Reinforcement learning is sometimes used for decision-making in autonomous systems like unmanned airborne vehicles, autonomous land vehicles, and guided weapons. Even manned systems use reinforcement learning-based automation to control imaging sensors for search and track capabilities, to maintain low level altitude, and avoid collision.

There are too many big data analytics methods to create a comprehensive list, but figure 1 presents an abbreviated list of the key methods discussed in this article, with brief descriptions. The choice of method will depend on the type of question that users ask, and the data that are available to answer this question. The methodological choices made by the user are essential to the effectiveness of big data, as they will affect the accuracy of the inferences drawn from big data. From this perspective, the effective use of big data requires human judgment.

[41] This paragraph is largely based on: Brandon Rohrer, 'Five Questions Data Science Answers', 31 Dec. 2015, https://brohrer.github.io/five_questions_data_science_answers.html.

<INSERT TABLE 1>

The Value of Big Data in National Security Contexts: The Intelligence Process

This section explores the value of big data and their use across a number of core intelligence activities. When doing so, we use the specific case of national security intelligence to answer broader questions and concerns about the ability to make effective uses of big data tools to advance the discovery and analysis of trends and threats.[42] At the most basic level, massive datasets enable higher confidence levels in inferring trends, patterns and anomalies. This capability can, in turn be harnessed through analytical methods to provide new informational outputs to national security professionals and the decision-makers they serve.

Big data pose notable challenges to the intelligence process. When massive amounts of data are amassed, stored, processed and used simultaneously, the intelligence process is compressed and inherently human activities, like analysis, risk being undermined by an overreliance on automation.[43] The reliance of intelligence agencies on big data tools further challenges traditional understandings of intelligence as a cycle that starts with specific intelligence requirement, then collection to fill knowledge gaps, analysis of the collected information, dissemination of the intelligence product or report, consumption by decision-makers and feedback. To avoid making limiting assumptions about intelligence as a process, we simply explore six core intelligence activities: requirements, collection, collation, analysis,

[42] See: McNeely and -Hahm, 'The Big (Data) Bang', p. 307; Honavar, 'The Promise and Potential of Big Data', p. 327.
[43] Michael Warner, 'The past and future of the Intelligence Cycle', in Mark Phythian, ed., *Understanding the Intelligence Cycle* (London and New York: Routledge, 2013), pp. 16–17.

dissemination, and security.[44] Given our focus on data and information, we decided not to discuss covert action, or operations that seek to 'influence the world by unseen means'.[45] The intelligence activities we selected are at the core of the national security decision-making process. When exploring the uses of big data in these activities, we find that big data can expand and improve core intelligence capabilities but not alter the human nature of intelligence.

Requirements

In the traditional model of the intelligence cycle, policymakers express intelligence requirements to intelligence managers who direct intelligence collection accordingly. In practice, as former CIA officer Arthur Hulnick notes, requirements are often 'derived internally within the intelligence system'.[46] Big data analytics' ability to discern general trends and anomalies in very large datasets – through anomaly detection and association algorithms – can help identify potential intelligence targets, thus driving intelligence requirements. The NSA, for instance, engages in bulk data collection, and uses data analytics capabilities to identify phenomena or targets of interest out of multiple large datasets and refine its collection effort. When doing so, the two first steps of the intelligence cycle model are inverted, and the collection and processing of a very wide range of potentially relevant data drives targeting or requirements.

Technological progress in this domain has allowed decision-makers to expect more from intelligence, thus posing new big data challenges. During a congressional testimony, then

[44] Peter Gill and Mark Phythian, 'From Intelligence Cycle to web of intelligence. Complexity and the conceptualization of intelligence', in Mark Phythian, ed., *Understanding the Intelligence Cycle* (London and New York: Routledge, 2013), p. 31; Arthur S. Hulnick, 'What's Wrong With the Intelligence Cycle', *Intelligence and National Security* 21: 6, 2006, pp. 959–979.
[45] Richard Aldrich, *The Hidden Hand: Britain, America and Cold War Secret Intelligence* (London: John Murray 2001), p. 5.
[46] Arthur Hulncik, 'Intelligence theory. Seeking better models', in Mark Phythian, ed., *Understanding the Intelligence Cycle* (London and New York: Routledge, 2013), p. 152.

Director of DARPA Regina Dugan explained that detecting a fighter over an area the size of Baghdad would require 100,000 times more data than detecting a strategic bomber in an area of the size of Reagan National Airport. The intelligence, surveillance and reconnaissance requirements to fight insurgents over wide areas require an ability to deal with data volume and variety in limited periods of time. In the words of the DARPA director 'The trend set by the detection of increasingly difficult targets is driving data volume exponentially'.[47]

Collection

Intelligence collection seeks to unearth new data and information to fill knowledge gaps. The improvement and exponential growth of intelligence collection capabilities in the last decades has provided the US intelligence community with increasingly large amounts of diverse data. Big data capabilities play an important role in this domain, facilitating the collection of massive amounts of data through indexing mechanisms, and data summarization algorithms, which automatically identify, summarize and store relevant data. Among different types of collection methods, open source intelligence provides tremendous potential for big data collection. This type of intelligence relies on data available in the pubic domain, a space where much information is easily accessible. Data collection tools automatically crawl through vast amounts of diverse data stored on the servers at the basis of the Internet, for example.

One of the traditional problems of open source collection is that it is subject to disinformation and propaganda. To overcome this problem, one expert notes, 'intelligence collectors have to develop screening algorithms to sort out what might be valuable'.[48] The

[47] US House of Representatives, Armed Services Committee, Subcommittee on Emerging Threats and Capabilities Statement by Dr. Regina E. Dugan (DARPA), 1 March 2011, pp. 10–11.
[48] Hulnick, 'Intelligence theory. Seeking better models', p. 153.

growth of fake news on the Internet, for example, has encouraged the development of new data analytics tools to extract information from online news articles in more reliable ways.[49] Clustering and classification algorithms can be leveraged to isolate fake news from the legitimate ones. The challenge is to identify the features that distinguish fake from legitimate news. The reputation of a news source and the content of articles are sometimes difficult to identify and analyse, and commonly used algorithms have not incorporated these features yet. Human intervention is still needed to identify the discriminative features of fake news and adapt algorithms accordingly. This is the model followed by companies like Facebook, which use algorithms to identify possibly fake news stories, and share these stories with a group of fact-checkers who assess their veracity and help identify fake news sites. The most obvious cases of fake news are down-ranked or even banished from Facebook feeds.[50]

The use of big data implies a willingness and capability to conduct bulk collection, which from a strictly practical point of view risks overwhelming the intelligence process with too much data to process and analyse. This emphasizes the role of collection managers in devising plans that can focus their resources on key threats.[51] In sum, human judgment continues to determine intelligence collection. In this context, big data do not alter the nature of intelligence but reinforces some of its traditional challenges such as identifying what to collect and what to discard.

[49] Deepa Seetharaman, 'Facebook Looks to Harness Artificial Intelligence to Weed Out Fake News', *Wall Street Journal*, 1 Dec. 2016.
[50] Josh Constine, 'Facebook now flags and down-ranks fake news with help from outside fact checkers', 15 Dec. 2016, https://techcrunch.com/2016/12/15/facebook-now-flags-and-down-ranks-fake-news-with-help-from-outside-fact-checkers/.
[51] Gary and Borum, 'Evolving Cyber Intelligence', p. 127.

Processing

The main contribution of big data to current intelligence practices is in the domain of processing and exploitation. Processing and exploitation turn raw data into usable information. A raw telecommunication intercept from a Russian target first needs to be processed from digital signals into symbols and text, and then translated (exploitation) to become usable information. Computers store and process large amounts of diverse data collected from multiple sources. For instance, surveillance drones like the MQ-1 Predator as well as many other signals and measurement and signature intelligence sensors are all transmitted, processed and exploited at core sites of the US Army Distributed Common Ground System.[52]

In practice, modern collection platforms both collect and process data. The Taranis, a drone produced by the company BAE Systems flies to preselected areas where its sensors capture multiple types of data, allowing its processor to identify a threat, an insurgent for instance, based on target behaviour, and alert human operators.[53] This capability is central to the discipline of activity-based intelligence, which 'integrates data from multiple sources around the interactions of people, events and activities, in order to discover relevant patterns, determine and identify change, and characterize those patterns to drive collection'.[54] Beside specific threat actors, data analytics can also use pattern recognition to detect trends in multiple large datasets that could, for example, point to growing instability in a specific region of the world, therefore

[52] Chandler P. Atwood, 'Activity-Based Intelligence: Revolutionizing Military Intelligence Analysis', *Joint Force Quarterly* 77: 2, 2015, p. 25.
[53] Bernard Marr, 'How AI, Drones and Big Data Are Reshaping The Future of Warfare', *Forbes*, 6 Oct. 2016, http://www.forbes.com/sites/bernardmarr/2016/10/06/how-ai-drones-and-big-data-are-reshaping-the-future-of-warfare/#9b3330b5bfe8.
[54] Atwood, 'Activity-Based Intelligence', p. 26.

improving situational awareness. Both the US Department of Defense and the IARPA have invested in such capabilities.[55]

The diversity of the big data streams creates a need for structure. Big data programs can automatically bring structure to the unstructured data found on websites for example, creating datasets that can 'talk' to each other, and allowing machines and humans to draw correlations across them. Natural language processing (NLP) is a field of computer science that is concerned with the ability of computers to automatically parse and process human languages. Documents released by Edward Snowden show how the NSA relies on data analytics to automatically transcribe audio conversations and make them searchable through key words.[56] Other NLP applications can be used to conduct sentiment analysis using social media feeds. An analysis of the sentiments expressed on Twitter in a specific region of the world can serve as an indicator for regime stability in specific countries, or even as a way to assess the impact of specific policies on target populations.[57] However, this sort of sentiment analysis is limited because not everyone uses Twitter or is able to tweet during a hurricane or a political upheaval. At the height of the Egyptian revolution of 2011, the government cut off nearly all access to the Internet and shut down mobile phone service, causing a '90 percent drop in data traffic to and from Egypt'.[58] Big data tools can process vast amounts of data, but the information drawn from these data is necessarily limited. Big data never is all the data.

---

[55] See Federal Business Opportunities, 'Information Volume & Velocity (IV2)', 13 Dec. 2012, https://www.fbo.gov/index?s=opportunity&mode=form&id=6fda262f46fab5f5273c18b1607e079d&tab=core&_cview=0; Intelligence Advanced Research Projects Agency, 'Open Source Indicators (OSI)', at https://www.iarpa.gov/index.php/research-programs/osi.

[56] National Security Agency, Human Language and Technology, 'For Media Mining, the Future is Now! (conclusion)', 7 Aug. 2006, https://theintercept.com/document/2015/05/05/media-mining-future-now-conclusion/; Office of the Director of National Intelligence, 'IARPA Announces Speech Recognition Challenge', ODNI News Release No.45-14, 18 Nov. 2014.

[57] Lim, 'Big Data and Strategic Intelligence', 629-630; Laurie A. Schintler and Rajendra Kulkarni, 'Big Data for Policy Analysis: The Good, The Bad, and the Ugly', *Review of Policy Research* 31: 4, 2014, p. 345.

[58] Matt Ritchel, 'Egypt Halts Most Internet and Cell Service, and Scale of Shutdown Surprises Experts', *New York Times*, 29 Jan. 2011, p. 13.

Big data applications facilitate analysts' access to large amounts of data, sometimes generated in real-time, through visual means. One example is the software Geofeedia, and some of its equivalents used across the US intelligence community. Geofeedia is an intelligence platform that gives analysts access to social media content in real-time based on their location. Geofeedia allows analysts to zoom on a city or neighbourhood within a specific timeframe, and get direct access to Twitter, Instagram and YouTube content posted by users located in that specific area.[59] This software relies on big data analytics to process vast amounts of diverse data and make them easily accessible to the analyst. It is up to the analyst to use and contextualize the outputs of this big data tool.

Analysis

Intelligence analysis is the 'thinking part' of the intelligence process.[60] Analysis can be defined as the application of knowledge, reasoning, and methods to transform raw data and information collected from multiple sources into informational outputs that are useful for decision-making. These outputs take the form of descriptions, forecasts, and explanations.[61] Descriptions address the 'what is' of a phenomenon of interest and would answer a question such as 'What are Russia's strategic goals in Eastern Europe?' Forecasts are prospective; they address issues and trends in the future and would answer a question such as 'Given Russia's previous activity in Ukraine, will it conduct a conventional land invasion in the next decade?' Explanations delve into causal mechanisms driving relationships and trends. They answer 'why' questions, such as 'Why does Russia use a hybrid warfare strategy against Ukraine and NATO?' These three

[59] Geofeedia, 'How it works', at https://geofeedia.com/products/how-it-works/; Conversation with a national security professional, 2016.
[60] George, Roger Z., and James B. Bruce, eds. *Analyzing Intelligence: National Security Practitioners' Perspectives*. Georgetown University Press, 2014, p. 3.
[61] The typology presented here is based on George and Bruce's judgments, forecasts, and insights.

outputs are nonexclusive; a forecast can incorporate an explanation of the causal forces leading to an event in the future.

Big data analytics generally focuses on correlations and, as such, is best used to help answer who, what, where and when types of questions.[62] In his comprehensive study of the role of big data in strategic intelligence, Kevjn Lim concludes that big data can help analysts 'discern long-term development, generate intelligence hypotheses, and adduce refuting facts'.[63] The ability of data analytics software to process vast amounts of data makes them ideal to identify trends and items of interest in large datasets. Data analytics methods like anomaly detection association and link analysis can be used to anticipate threats, for example, to identify possible targets of radicalization.[64] Beside the identification of threats and targets, data mining programs can detect general patterns of behaviour among target populations and more specific patterns in near real time that point to threats or phenomena of interest.[65] Online trends can, for instance, serve as indicators of offline events. During the first two weeks of the 2011 Egyptian revolution, 'over 32,00 new groups and 14,000 new pages were created on Facebook in Egypt'.[66] Monitoring Twitter traffic and content through big data tools can similarly help identify emerging events as they occur.[67] According to Andrew Hallman, the deputy director for digital innovation at the CIA, his agency was able to improve 'forecast to the point of being able to anticipate the development of social unrest and societal instability to within three to five days out'.[68] Overall,

[62] On big data and correlation see: Lim, 'Big Data and Strategic Intelligence', p. 626.
[63] Lim, 'Big Data and Strategic Intelligence', p. 619.
[64] Neil Couch and Bill Robins, 'Big Data for Defence and Security', *RUSI Occasional Paper*, 2013, p. 10.
[65] William J. Lahneman, 'IC Data Mining in the Post-Snowden Era', *International Journal of Intelligence and* Counterintelligence 29: 4, 2016, pp. 714–15.
[66] Robert E. Wilson, Samuel D. Goslin, and Lindsay T. Graham, 'A Review of Facebook Research in the Social Sciences', *Perspectives on Psychological Science* 7: 3, 2012, p. 27.
[67] Jamie Bartlett and Carl Miller, *The State of The Art: A Literature Review of Social Media Intelligence Capabilities for Counter-Terrorism* (London: Demos, 2013), 28.
[68] Frank Konkel, 'CIA can anticipate social unrest 'three to five days' out in some cases', 4 Oct. 2016, http://www.nextgov.com/defense/2016/10/cia-can-predict-social-unrest-three-five-days-out/132102/.

big data tools can help analysts describe, sometimes even identify, and forecast situations based on a wide array of data sources. When doing so, they facilitate the task of intelligence analysts but cannot replace them.

Data analytics can contribute to intelligence analysis but cannot substitute it because analysis is a human activity that requires judgment and contextualization. Data mining software such as those used in the NSA PRISM program sift through vast amounts of telecommunications metadata to identify patterns or correlations between different variables of interest, suggesting that a specific individual or group of individuals might be a threat. However, these software can only establish correlation. Human treatment of the subject remains necessary to assess the trends and red flags identified by automated computer systems, and seek a search warrant to refine the collection effort when deemed necessary.[69]

The consumers of big data, in this case analysts, play an important role ensuring that the patterns emerging from data mining tools are relevant. To avoid seeing patterns where none actually exist, sociologists Patricia White and R. Saylor Breckenridge point out that 'it is crucial to begin asking questions about the analytic assumptions, methodological frameworks, and underlying biases embedded in the big data phenomenon'.[70] Analysts play a key role contextualizing the results of big data analytics processes and merging these results with small data or specific cases to produce an actionable, timely and comprehensive report. Data analytics tools help analysts process vast amounts of data to focus their effort on sense making rather than on processing raw data.[71] However, big data are only but one of the tools in the analyst's toolbox.

[69] Lahneman, 'IC Data Mining in the Post-Snowden Era', pp. 704, 710.
[70] Patricia White and R. Saylor Breckenridge, 'Trade-Offs, Limitations, and Promises of Big Data in Social Science Research', *Review of Policy Research* 31: 4, 2014, p. 336.
[71] Couch and Robins, 'Big Data for Defence and Security', pp. 9, 11.

Dissemination

Big data tools can help disseminate intelligence from producers to consumers. Beside the standardized routes used in most intelligence systems to disseminate periodic and occasional reports, data analytics tools can help analysts convey information more effectively and faster. One example might be applications such as the recommendation engine or 'you might like' feature currently available on e-commerce sites like Amazon or newspapers websites. A recent report produced by Gregory Treverton suggests that similar applications are used in the US intelligence community. Feeds on the interagency micro-blogging tool eChirp, for example, 'provide notice of thought-provoking or special items'.[72] Similar applications might be used in the intelligence community to suggest (or push) relevant intelligence products to specific consumers, be they analysts or decision-makers. When doing so, big data expands and refines the dissemination of intelligence products.

Visualization tools relying on data analytics capabilities can help humans understand complex realities based on vast amounts of diverse data sources, and facilitate well-informed decision-making. The Carter Center, for instance, has developed an interactive map that tracks the evolving frontlines of the current conflict in Syria, using open source data processed through a software developed by Palantir, which is widely used in the US intelligence community.[73] This map documents over 70,000 conflict events in Syria and shows the changing relations between multiple armed groups and movements of displaced people across time. In this case, a big data application represents a complex reality in an accessible and interactive way. Analysts can use similar products to tailor their briefing in real-time and best answer the need of their consumer.

[72] Gregory F. Treverton, *New Tools for Collaboration. The Experience of the U.S. Intelligence Community* (New York: Rowman & Littlefield, 2016), p. 13.
[73] The Carter Center, 'Tracking the Front Lines in Syria', https://d3svb6mundity5.cloudfront.net/dashboard/index.html.

Consumers could also use the product on their own, developing their situational awareness of constantly evolving threats on demand, and possibly without an analyst.[74] From this perspective, the appeal of big data visualization tools might not only augment, but also threaten the traditional role analysts have played in briefing intelligence to decision-makers.

Beside maps, visualization tools can facilitate human understanding of complex phenomena, showing recognizable patterns and trends in vast networks.[75] One of the most well researched applications of this capability is the visualization of terrorism data.[76] Here the results of big data analytics can show linkages between various terrorist organizations, subgroups within larger terrorist networks, and individuals who are at the centre of specific networks of militants. Such capabilities can be used to understand and monitor the flow of information and individual connections within specific communities, providing warning or actionable intelligence that, when put into context, can help intelligence and security services to disrupt terrorist organizations.[77]

Counterintelligence and security

In a narrow sense counterintelligence and security aim to protect intelligence agencies against penetration by adversary services. A broader and more common understanding of counterintelligence and security encompasses defence against major threats to national security,

[74] The Carter Center, 'Syria Conflict Resolution', https://www.cartercenter.org/peace/conflict_resolution/syria-conflict-resolution.html.
[75] M. Shahriar Hossain, Patrick Butler, Arnold P. Boedihardjo, and Naren Ramakrishnan, 'Storytelling in Entity Networks to Support Intelligence Analysts', ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12), 2012; M. Shahriar Hossain, Christopher Andrews, Naren Ramakrishnan, and Chris North, 'Helping Intelligence Analysts Make Connections', AAAI Workshop on Scalable Integration of Analytics and Visualization, 2011, pp. 22–31.
[76] Lavanya Venkatagiri Hegde, Nerella Sreelakshmi and Kavi Mahesh, 'Visual Analytics of Terrorism Data', IEEE International Conference on Cloud Computing in Emerging Markets, 2016, pp. 90–94.
[77] Sean F. Everton, *Disrupting works* (Cambridge: Cambridge University Press, 2012); Bartlett and Miller, *The State of the Art*, pp. 35–44.

including espionage, but also terrorism and transnational crime. One security application of big data analytics, specifically through NLP capabilities, is the identification of malicious domains and malicious codes (malware) in cyberspace.[78] Automated data analytics can be used as a part of broader systems to defend computer networks. In the field of cyber-security, network based intrusion detection systems monitor Internet traffic, looking for specific signatures or codes that deviate from the norm or have already been identified as malware.[79] Such systems help analysts spot advanced persistent threats and automatically block cyber attacks. Cyber attacks take place at the speed of light, and this raises interesting questions about the diminishing role of humans in national security decision-making. When network intrusion detection systems analyse vast amounts of data to automatically block cyber threats, big data analytics effectively replaces humans. Yet big data capabilities are not a panacea and the inability of algorithms to take into account the broader context of an attack can make it hard for machines to detect social engineering scams on their own.[80]

Conclusion: big data are not always better

To date, security studies researchers have not explicitly defined or instituted a framework for assessing the big data phenomenon. To fill this gap in the literature, we explored the characteristics, technology and methods of big data and situated them in the context of national security. Our exploration of big data in traditional intelligence activities – requirements,

[78] Bobby Filar, 'NLP For Security: Malicious Language Processing', 19 Aug. 2015, at https://www.endgame.com/blog/nlp-security-malicious-language-processing.
[79] Edward Skoudis, 'Information Security Issues in Cyberspace', in Franklin Kramer, Stuart Starr, Larry Wentz, eds, *Cyberpower and National Security* (Washington DC: Potomac Books Inc., 2009), 190-191; Paul Giura and Wei Wang, 'Using Large Scale Distributed Computing to Unveil Advanced Persistent Threats', *Science Journal* 1: 3, 2012, pp. 93–94.
[80] Erik Gartzke and Jon R. Lindsay, 'Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace', *Security Studies* 24:2, 2015, p. 333

collection, processing, analysis, dissemination, and counterintelligence and security – suggests that technological advances have allowed security professionals to collect and process larger and more diverse amounts of data, sometimes rapidly, so that they can be analysed and intelligence can be disseminated more effectively. These strengths, and the limitations of traditional intelligence disciplines like HUMINT, explain why big data tools have played an increasingly prominent role in national security processes.

However big data are not always better than humans. Beginning in the 1950s, scholars in the field of psychology discussed the validity of judgments made by professionals, such as doctors, versus the outputs of actuarial formulas, a class of algorithm based on pre-specified input from experts. For example, a clinical actuarial formula for diagnosing illness would take into account the types of patient symptoms (runny nose, but no headache) and weight each to make a judgment (diagnosis is allergies, not flu). A key figure in the debate, Paul Meehl, concluded that in some tasks algorithms are superior and in others human judgment is necessary. To make this point Meehl used two examples.[81] He first asked readers to imagine they have a basket full of groceries at the checkout and asks what would be the best way to accurately total the bill? Would it be: 1) have the cash register add up the cost of the groceries or 2) let the cashier make an estimated guess? The first choice is the correct answer; the cash register's machine does a better job summing up the bill than the cashier's brain. In his a second example, Meehl invites his readers to predict whether a professor will see a specific movie today. The proposed algorithm 'employs factors such as the day of the week and the type of movie available to make its prediction. However, the prediction fails because the algorithm cannot consider the fact that the professor cannot leave the house because he has a broken leg, a condition that

[81] Paul Meehl, 'Causes and effects of my disturbing little book'. *Journal of personality assessment* 50: 3, 1986, pp. 370–375.

negates the normally accurate forecast made by the algorithm'.[82] This second case demonstrates that even if one has 'all the information…about the [professor's] predilections', making a wrong prediction is still possible.[83]

The lesson is that when environments are predictable, such as inventorying prices in a grocery store, algorithms will almost always outperform human judgment. However, in unpredictable environments characterized by sudden, dramatic changes, automated analysis is likely to be wrong.[84] The security environment is characterized both by long-term trends, which are most visible at the strategic level, and sudden, dramatic changes causing surprises in the short term. In the latter situations experts – who can follow their intuition and think outside the box – are essential to take into account 'broken-leg' variables. Research in the field of forecasting reinforces this lesson and finds that human judgment combined with algorithms are significantly more accurate than algorithms or human judgment alone.[85]

When used on their own machines and the deterministic algorithms they use 'strip out much of the context' in which humans interact and are 'oblivious to social clues or shades of agreement'.[86] Some important national security insights, such information on the intentions of foreign leaders, are not easily expressed through data. An important realization then is that big data cannot, nor should, replace the central role of humans, be they producers or consumers of intelligence, in national security decision-making. Big data applications are best used when they free humans 'to do what they do well - think, ask questions, and make judgments about complex

[82] Hal Arkes R. and James Kajdasz, 'Intuitive Theories of Behavior', in Baruch Fischhoff and Cherie Chauvin, eds., *Intelligence Analysis: Behavioral and Social Scientific Foundations*, (Washington, DC: National Academies Press, 2011), p. 152.
[83] Kurt Salzinger, 'Clinical, Statistical, and Broken-leg Predictions,' *Behavior and philosophy* 33, 2005, p. 93.
[84] Paul Meehl, *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence* (Northvale, NJ: Jason Aronson, 1996); Peters, J. T., Hammond, K. R., & Summers, D. A., 'A note on intuitive vs. analytic thinking', *Organizational Behavior and Human Decision Processes* 12: 2, 1974, pp. 125–131.
[85] Matthias Seifert and Allegre Hadida, 'On the relative importance of linear model and human judge(s) in combined forecasting', *Organizational Behavior and Human Decision Processes* 120, 2013, pp. 24–36
[86] Gartzke and Lindsay, 'Weaving Tangled Webs', p. 333.

situations'.[87] In the age of big data, as Cukier and Mayer-Schoenberger note, 'the most human traits will need to be fostered—creativity, intuition, and intellectual ambition—since human ingenuity is the source of progress'.[88] These human characteristics can help refine the intelligence process by taking into account unexpected variables like the 'broken leg', or even discard false positive errors.[89] The future of big data and national security lies in humans' ability to embrace the power and mitigate the limits of algorithms. Doing so requires a better understanding of the role big data is playing in core national security functions like intelligence.

[87] US Department of Homeland Security, Enabling Distributed Security in Cyberspace – Building a Healthy and Resilient Cyber Ecosystem with Automated Collective Action, 23 March 2011, Washington DC, p. 8.
[88] Cukier and Mayer-Schoenberger, 'The Rise of Big Data', p. 40.
[89] For a good example, see: Brantly, 'Changing the Game', 144.

Data Analytics Method	Objective
Anomaly detection	Identifying items, events, or observations that do not conform to an expected behaviour or pattern. The definition of anomaly varies from dataset to dataset resulting in a wide variety of algorithms designed to process different data types including time-series, stream, network, text, video, and imagery.
Association	Discovering interesting relationships hidden in large datasets. These relationships are generally retrieved from frequent usage patterns.
Classification	Assigning objects in a collection of data to target categories or classes. The goal of classification is to accurately predict the target class for each sample in the data. A classification model could be used to identify an intercepted phone call as a part of zero, low, medium, or high-risk suspicious activity.
Clustering	Grouping a set of objects or data points. Clustering brings similar objects together in the same group. The notion of similarity depends on the specific analytic task.
Link analysis	Defining, discovering, and evaluating relationships between objects or data points. Data points may be nodes in a graph or network connecting people, organization, or other entities.
Recommendation	Filtering to produce a narrow ranked list of resources for a specific task within a particular context. Recommendation systems can help decision-making by providing suggestions regarding agency, personnel and expertise in emergency situations relevant to a specific threat-context.

Table 1. Select Data Analytics Methods