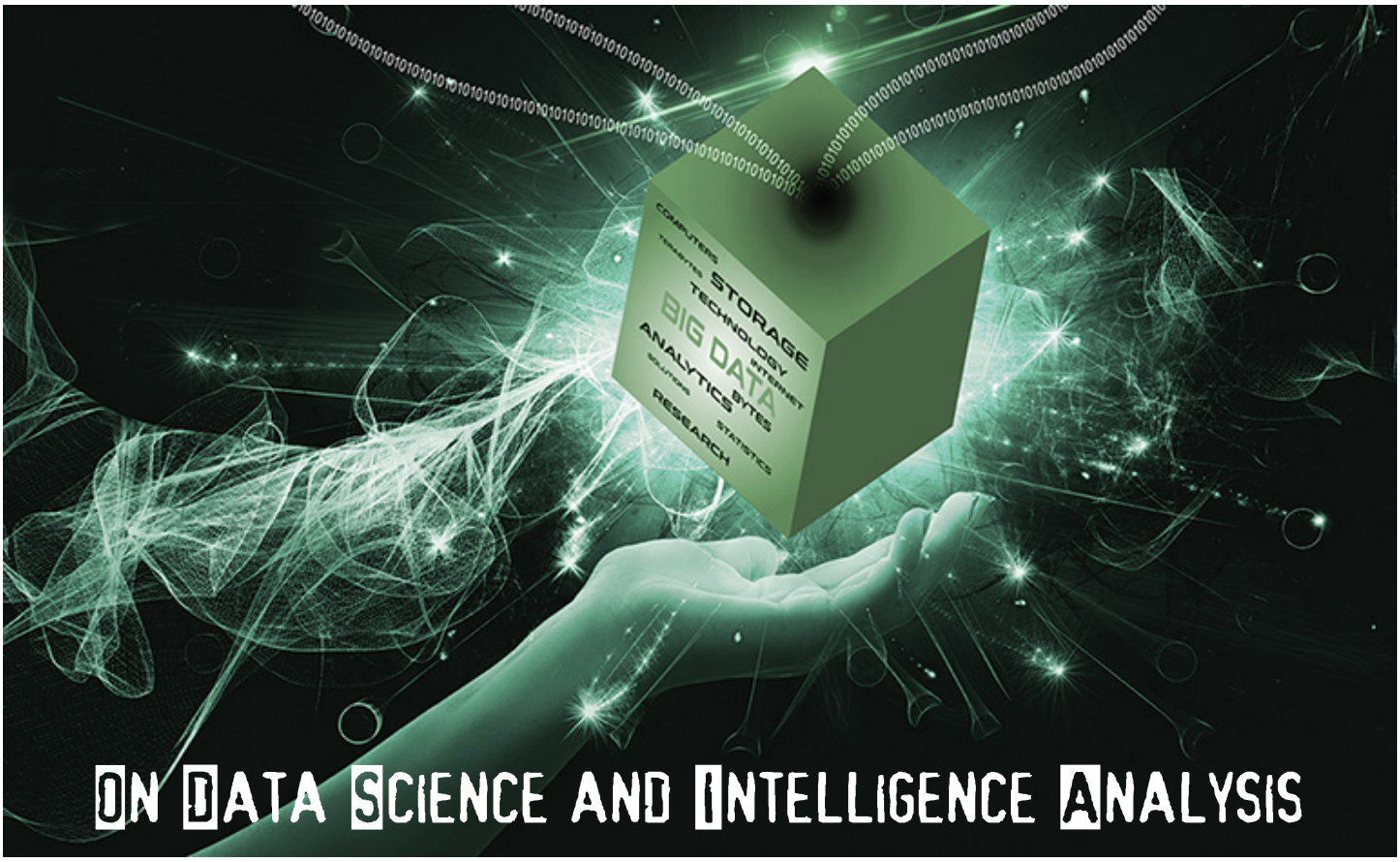

On Data Science and Intelligence Analysis

by Captain Iain J. Cruickshank

Editor's Note: This article is the first in a two-part series on data science. This first part provides a basic foundational understanding of data science and its application in the intelligence community. Part two of the series, by CPT Jason Boslaugh and Mr. Zachary Kendrick, will be published in the October-December 2019 issue. Their article discusses how the U.S. Army can apply data science lessons learned from academia and industry to modernize the intelligence warfighting function.

Introduction

There has been much hype in recent years about data science and big data—some well justified and some unwarranted. Many people see data science, and its associated disciplines of machine learning and artificial intelligence, as a panacea for the ills of modern decision making and analysis, especially those ills that plague the intelligence community.[1] However, some members of the intelligence community completely dismiss how military intelligence analysis can benefit from data science and machine learning. In this article, I will—

- clarify some of the roles of data science in intelligence analysis,
- describe conditions for its successful use, and
- promote an environment that maximizes the use of modern data science for better intelligence analysis.

Data science is a "multi-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from data in various forms, both structured and unstructured."[2] In simpler terms, it is the ability to distill knowledge (i.e., useful information to humans) from data (i.e., raw text, images, signals, etc.), typically using a computer. Data science often employs machine learning and other artificial intelligence techniques. As such, it is the preferred discipline for analyzing big data, online social networks, and other data sources that are simply too large, heterogeneous, and dynamic for any single human to comprehend.

Despite all the hype, data science is not a replacement for intelligence analysis. Intelligence analysis relies on much more than what data science as a discipline provides.[3] Intelligence analysis relies on human intuition and experience. Rather than replace analysts, data science transforms analysts' jobs in ways that make analysts more effective, focusing on applying human intuition. They are able to redirect the majority of their time and effort from the acquisition of information (i.e., reading reports or watching full motion video feeds) to thinking about the adversary, the environment, and trends in the battlespace.

Why Data Science?

As documented in a spate of recent articles, a trend has emerged to use more open-source and social network information for intelligence analysis.[4] These sources of information can be invaluable but come in a deluge of constantly changing, error-prone data. Despite this, more intelligence analysts are relying on this type of data because it allows for insight that analysts cannot otherwise obtain by looking at other sources of information.[5] Recent articles from the military intelligence community highlight several problems, including—

- A lack of interoperable and integrated data sources, and a lack of a common operational picture and ontology needed to understand information.[6]
- A lack of methods that can make information intuitive to understand and provide patterns easy to comprehend for an analyst. Furthermore, a lack of having methods that can adapt to different warfighting domains (i.e., cyberspace, counterinsurgency, etc.).[7]
- Overwhelming and contradictory reporting, as well as filtering of available information for an analyst.[8]
- A large amount of time spent monitoring full motion video feeds or sifting through volumes of reports to find a few actionable pieces of information.

These problems coalesce around intelligence analysts being able to get the "right" information and have the information presented in a way that enables their analyses. While analysts now use more and varied information for intelligence analysis, issues persist with how to process that voluminous, varied information to allow analysts to exploit it. In this situation, I would argue that data science flourishes as a discipline.

All the recent developments in industry by companies like Google, Facebook, and Microsoft have boosted interest in data science and artificial intelligence, and the Department of Defense has taken notice. The 2018 National Defense Strategy and its service-derivative documents all prominently feature artificial intelligence and modern, digital technologies.[9] Recent advances in the Internet of things and drones have led to breakthroughs in collection technology for intelligence purposes.[10] Members of the military intelligence community have also pushed to incorporate the tools of data scientists, most notably machine learning.[11] What these articles lack, however, is guidance on how we should implement data science and machine learning for intelligence analysis.

But there is a caveat for the use of data science and artificial intelligence—*simply buying off-the-shelf machine learning and artificial intelligence products will be insufficient.* Here's why:

First, data science is still a new discipline. As a new discipline, many of its tools and methods are not engineered for people outside of specific applications, like the technology industry and academia. If analysts cannot understand the tools, analysts will not use them.

Second, no unique algorithms or models exist that will apply in all situations. The more complex and dynamic the data environment, the less any one given algorithm or set of tools will be applicable.

Third, off-the-shelf tools are not necessarily the solution. With the rise of adversarial machine learning, and the fact that the enemy always gets a vote in any conflict, any off-the-shelf tool will need to be adapted and changed as the enemy develops its algorithmic countermeasures. A great example of what adversarial machine learning can do are the recent studies on fooling image detection machine learning algorithms.[12] And if the tool is not open source, but rather some proprietary product, adapting it to the realities on the battlefield may be impossible. Thus, while it may be attractive to buy off-the-shelf artificial intelligence and machine learning tools, these tools will not meet the unique needs of intelligence analysis.

The Relationship

Data science can enable intelligence analysis by its ability to digest large, heterogeneous, error-prone data into human-usable information like trends, outliers, and key data points. To better illustrate this point, I will begin with a simple analogy—gold mining. The whole point of the gold-mining industry is to find gold and mine it from the earth. This gold is critical for various uses around the world. Now, a gold miner can only dig up so much earth; if he digs in the wrong places, he will fail to produce gold. A land surveyor, on the other hand, can find likely places in the earth where there is gold, but the land surveyor is not skilled in mining. Thus, a beneficial arrangement arises. The land surveyor indicates where gold is likely to exist, and the gold miner skillfully extracts it from the earth. The land surveyor's expertise leads to the gold miner extracting more gold because he is only digging where the gold is likely to be. Now, substitute the terms *intelligence analyst* for *gold miner, data scientist* for *land surveyor,* and *actionable intelligence that drives commanders' decisions* for *gold,* and you will understand how data science can significantly empower intelligence analysis. An intelligence analyst applies meaning and forms intelligence estimates based on the trends and patterns that a data scientist was able to surface.

Without diving into specific algorithms and methods, the following are some common ways data science can enable intelligence analysis:

- Allows entity detection on full motion video (i.e., Project Maven), so that analysts do not have to keep their eyes glued to just one full motion video screen for endless hours, but rather are cued to the video when something of interest happens.
- Characterizes reports, images, and other forms of information into more visual and interpretative formats for quick understanding of the intelligence environment.
- Provides predictive analytics that are constantly running, based upon key indicators of some event of interest.
- Expresses uncertainty in information available in both the information space and geographically.
- Identifies anomalies in information and in target behavior.
- Provides content recommendation to help identify more pieces of relevant information based on what analysts identify as relevant information for their intelligence analyses.

While this is not a complete list, it is important to note that at no point does data science replace intelligence analysis. None of these methods explains why an adversary might do what they do, nor do these methods negate analysts' judgments. Rather, they are tools that allow analysts to focus their effort on thinking about those key pieces of information and making informed judgments, instead of searching for information or detecting anomalies. Furthermore, removing extraneous information will allow analysts not only to spend more of their effort on actual analysis, but also to produce better intelligence.[13]

One final point about the relationship between an intelligence analyst and a data scientist. Data science uses "feature selection," which is the process of selecting variables from the information you will use in your models. A data scientist uses variables present in the data that actually relate to the phenomenon that is being analyzed. Subject-matter expertise is tremendously important in this area, and thus, it is where the intelligence analyst must play a role. When it comes to selecting attributes about the enemy, populace, or terrain, an intelligence analyst needs to be involved in that selection. So, when it comes to feature selection, the features should be a combination of what an intelligence analyst deems important and what a data scientist can meaningfully use.

Enabling Data Science for Intelligence Analysis

Data scientists occasionally say that for good data science, they need three elements:

- a good problem,
- computational resources, and
- available data.

Intelligence analysis presents the needed "good problems," but intelligence systems are generally not set up for proper computational resources and data management.

A Good Problem. In the field of intelligence analysis, there are many good problems for data scientists to resolve. What is more, as data science is incorporated into intelligence analysis, many more good problems are likely to arise.

Computational Resources. One resource that all data science methods and algorithms require is computing power. Typically, most industry and academic data scientists rely on cloud computing for this power. Cloud computing allows data scientists to create a virtual machine, from anywhere in the world, that has the exact specifications they need for the data science analyses they are going to perform, with far more computational power than can realistically be carried about. Cloud computing also enables economies of scale for computing resources, which is critical for any large-scale organization. The need for cloud computing resources is well known to the intelligence community. The community has, however, consistently failed to materialize this cloud infrastructure for a variety of reasons.[14] New attention and urgency should be paid to getting a cloud architecture working to enable data science for intelligence.

There are also considerations about how data science will work in austere and electromagnetically degraded environments. For example, virtualization through things like VMware or Docker provides cheap and effective ways for data scientists to continue to use their tools on local machines, without needing additional equipment or a connection to the wider enterprise. All that is required is the ability to have virtual machines on those physical machines with access to classified networks. Finally, it is important to note that many data science tools are open source and constantly updating, and they require a programming ability in the R or Python programming languages. Therefore, closed, static data intelligence suites like the Distributed Common Ground System-Army are not suitable. They do not allow for data scientists to program and bring in their own tools to the computational environment where the data lives.

Available Data. Data management is the other major issue hampering the use of data science for intelligence analysis.

As the intelligence community has already remarked, the varied, heterogenous databases of information make getting access to any information, nonetheless the right information, exceedingly difficult.[15] In essence, these problems result in having to manually search through intelligence databases or download by hand from intelligence analysis suites the information that is vital to data science. This manual labor will result in a huge loss of time for analyses and will likely be missing information. Therefore, it is critical for data science-informed intelligence analysis to have systems that have mandated programmatic access to their data from things like application programming interfaces and standard data ontologies.

Conclusion

Intelligence analysis is increasingly relying upon greater, more heterogeneous sources of information. As a result, the sheer amount of information is far exceeding what an intelligence analyst can parse while still producing actionable intelligence. Problems with inconsistent, dynamic, and erroneous information continue to plague intelligence analysis. All these problems naturally point to a solution—using data science in intelligence analysis, not as a replacement for intelligence analysts but as a means for enabling faster, better quality intelligence.

Some key conditions exist, however, like cloud computing and appropriate data management, which still must be addressed in intelligence systems in order to enable data science. Once addressed, data science will provide a distinct combat advantage to whichever force is best able to employ it as part of its intelligence analysis process. If the trends regarding greater volumes of digital information being available for intelligence continue into the future—as they almost certainly will—it is not unreasonable to expect intelligence analysts to take on many of the skills of data science, while data scientists work on core algorithmic development and specialty analyses. To get to this point, we need to establish a solid and beneficial working relationship between data scientists and intelligence analysts.

Endnotes

1. Drew Conway, "Data Science in the U.S. Intelligence Community," *IQT Quarterly* 2, no. 4 (Spring 2011): 24-27, https://static1.squarespace.com/static/5150aec6e4b0e340ec52710a/t/51525211e4b0e9fad0b56f9c/1364349457311/IQT-Quarterly_Spring-2011_Conway.pdf.

2. Vasant Dhar, "Data Science and Prediction," *Communications of the ACM* 56, no. 12 (December 2013): 64-73, https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext.

3. John Coyne, "The Future of Intelligence Analysis: Computers versus the Human Brain?" *The Strategist*, Australian Strategic Policy Institute Blog, 12 September 2017, https://aspistrategist.siteindev.com.au/future-intelligence-analysis-computers-versus-human-brain/.

4. Paul B. Symon and Arzan Tarapore, "Defense Intelligence Analysis in the Age of Big Data," *Joint Force Quarterly* 79, 4th Quarter (October 2015), https://ndupress.ndu.edu/Media/News/Article/621113/defense-intelligence-analysis-in-the-age-of-big-data/.

5. Damien Van Puyvelde, Shahriar Hossain, and Stephen Coulthart, "National security relies more and more on big data. Here's why," *Washington Post,* 27 September 2017, https://www.washingtonpost.com/news/monkey-cage/wp/2017/09/27/national-security-relies-more-and-more-on-big-data-heres-why/?utm_term=.97152b2b58e9.

6. Donald Beattie and Robert Coon, "Improving Intelligence Sharing," *Military Intelligence Professional Bulletin* 44, no. 4 (October-December 2018): 25-28.

7. Christie P. Cunningham, "Developing Analytic Capabilities in Changing Environments," *Military Intelligence Professional Bulletin* 44, no. 4 (October-December 2018): 68-71.

8. Alex Morrow and Michael Dompierre, "Getting Intelligence to Move at the Speed of Decisive Action," *Military Intelligence Professional Bulletin* 44, no. 4 (October-December 2018): 55-58.

9. Office of the Secretary of Defense, *Summary of the 2018 National Defense Strategy of the United States of America,* n.d., http://nssarchive.us/national-defense-strategy-2018/.

10. Daniel C. Tuttle and Robert D. Sensenig II, "Strength Begins with Science and Technology," *Military Intelligence Professional Bulletin* 44, no. 1 (January-March 2018): 25-30.

11. Robert Collins, Lindsay Yowell, and Greg Hartman, "The Future of Intelligence Analysis, Analytics, and Distribution," *Military Intelligence Professional Bulletin* 44, no. 4 (October-December 2018): 35-37.

12. Drew Harwell, "Amazon facial-identification software used by police falls short on tests for accuracy and bias, new research finds," *Washington Post,* 25 January 2019, https://www.washingtonpost.com/technology/2019/01/25/amazon-facial-identification-software-used-by-police-falls-short-tests-accuracy-bias-new-research-finds/?noredirect=on&utm_term=.4bb5999b6004.

13. Zeynep Tufekci, "Big Data and Small Decisions," *Scientific American* 320, no. 3 (March 2019).

14. Naomi Nix, "Inside the Nasty Battle to Stop Amazon From Winning the Pentagon's Cloud Contract," *Bloomberg,* December 20, 2018, https://www.bloomberg.com/news/features/2018-12-20/tech-giants-fight-over-10-billion-pentagon-cloud-contract.

15. Beattie and Coon, "Improving."

CPT Iain Cruickshank is currently a Ph.D. candidate in societal computing at Carnegie Mellon University as a National Science Foundation Graduate Research Fellow. His previous assignments include Company Commander for D Company, 781st Military Intelligence Battalion (Cyber), and sub-element lead for planning and analysis and production on a National Mission Team in the Cyber National Mission Force. CPT Cruickshank holds a bachelor of science from the U.S. Military Academy. He also holds a master of science in operations research from the University of Edinburgh, obtained as a Rotary Ambassadorial Fellow. He has deployed to Paktia, Afghanistan, as a member of Security Force Advise and Assist Team.