



DATA SCIENCE IN THE U.S. INTELLIGENCE COMMUNITY

By Drew Conway

As terms of art, “big data” and “data science” are relatively new. A search of Google Trends indicates that the term “big data” did not emerge in news references until mid-2007 and recently peaked in the fourth quarter of 2010. This brief surge may be explained by the parallel emergence of inexpensive storage technology and high quality open-source distributed computing platforms, such as Hadoop (Olson 2010). The term “data science,” however, has emerged even more recently. This has led to ambiguous and contradictory definitions, especially regarding the skills a data scientist actually needs. These skills are of particular interest to the U.S. Intelligence Community, and it is exactly this intersection where I would like to focus.

While the cost to set up a large capacity, high-performance, computing environment is still prohibitive for many small to medium firms, companies like Amazon (Harris 2009) and Google (Baker 2007) have established extensive infrastructure that allows anyone to use these tools at very low cost.

This has led to a democratization of big data, wherein anyone with a data set and the inclination to analyze it can use cutting edge open-source tools to do so with relative ease. For analysts and IT professionals in the U.S. Intelligence Community (IC), issues related to the storage, manipulation, and analysis of data on the

order of petabytes are not new. In fact, the community has been using and developing technology to handle this level of information scaling for many years.

Given this experience, and the informational advantage it affords the IC, what can the community learn from observing this industry as it matures through this period of technological adolescence? Currently, the focus within the industry is on the tools being developed to handle this scaling (Loukides 2010), but much more than a high-performance computing cluster is needed. Fundamentally, the data science movement is about how the people and the tools drive

innovation and promote discovery. This is what data science must be at its core, and for the Intelligence Community to benefit from this it must cultivate these principles within its analysts.

To begin, the term “data science” is a bit of a misnomer, but is useful when attempting to define the set of skills required to extract insight from large data. Science is about discovery, and as discussed above, up to now much of the focus in the data science and big data communities has been on the tools. The work of a data scientist in the IC, however, is only concerned with tools insofar as they are useful in answering a question.

It is difficult to narrowly define the skills of a data scientist because they are naturally interdisciplinary, yet they exist at intersections of disciplines that do not often merge. In a general sense, there are three primary areas of expertise needed to be a successful data scientist.

First, one must have hacking skills. By this I do not mean malicious computer hacking or unauthorized disclosure of information. Rather, hacking skills in this context mean proficiency working with large, unstructured chunks of electronic data. Put simply, a hacker is someone who can easily navigate the required set of tools needed to be a data scientist. Second, one needs a basic understanding of mathematics and statistics, as these fundamentals will inform all of the analysis. Finally, and perhaps most importantly, a data scientist must have some substantive expertise in the data being analyzed. As mentioned, for data science to be a science it must foremost be about discovery. Questions must drive analysts’ curiosity and motivate their work.

To understand how these three sets of skills intersect and why they are important to the U.S. Intelligence Community, I present the following Data Science Venn Diagram as an illustrative abstraction. Note, none of the pieces is discipline specific, but rather could be thought of as conglomerates of many areas of expertise. More importantly, each of these skills is valuable on its own, but when combined with only one other is at best simply not data science, or at worst downright dangerous. Data science occurs only at the intersection of the three primary skills.

Data is a commodity traded electronically; therefore, in order to be in this market you need to be competent using the technology. This, however, does not require a background in computer science. In fact, many of the most accomplished data scientists I have met never



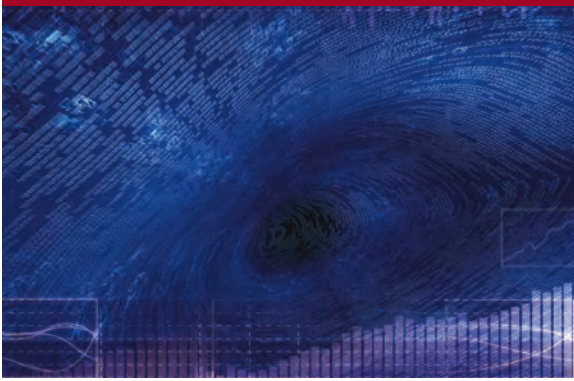
Figure 2 | The Data Science Venn Diagram

took a single CS course. Being able to manipulate text files at the command-line; understanding regular expression and UNIX tools like sed, awk, and grep; and thinking algorithmically — these are the skills that make a successful data hacker.

While much of the data available to the Intelligence Community is relatively well structured, and can include metadata about time, place, etc., a great deal of the most useful data is not. Raw data from the Internet or other means of collection are messy and require refinement. The data scientist’s time is first spent getting data into a workable format; therefore, without knowledge of how to do this, the work can never begin.

In addition, data is not always handed over to the analyst. An analyst’s job is to convey novel findings by synthesizing information from many different sources. The most useful information, however, may not come with a predesigned lever for acquisition. Understanding how to scrape the web, or parse data from various social media outlets and interact with web-based APIs, will become a more regular routine for data-driven intelligence analysis. For an analyst to begin to fulfill the role of a data scientist, a high degree of hacking skills will be required.

Once the analyst has acquired and cleaned the data, the next step is to actually extract insight from it. In order to do this, one must apply the appropriate methods of mathematics and statistics. This procedure requires at minimum a baseline familiarity with these tools. This is not to say that a Ph.D. in statistics is



Understanding how modeling assumptions impact the interpretations of analytical results is critical to data science, and this is particularly true in the IC.

required to be a competent data scientist, but it does require knowing the assumptions of a basic statistical model and their consequences for an analysis.

For example, it is common for an intelligence analyst to measure the relationship between two data sets as they pertain to some ongoing global event. Consider, therefore, in the recent case of the democratic revolution in Egypt that an analyst had been asked to determine the relationship between the volume of Twitter traffic related to the protests and the size of the crowds in Tahrir Square. Assuming the analyst had the data hacking skills to acquire the Twitter data, and some measure of crowd density in the square, the next step would be to decide how to model the relationship statistically.

One approach would be to use a simple linear regression to estimate how Tweets affect the number of protests, but would this be reasonable? Linear regression assumes an independent distribution of observations, which is violated by the nature of mining Twitter. Also, these events happen in both time (over the course of several hours) and space (the square), meaning there would be considerable time- and spatial-dependent bias in the sample. Understanding how modeling assumptions impact the interpretations of analytical results is critical to data science, and this is particularly true in the IC.

The third critical piece — substance — is where my thoughts on data science diverge from most of what has already been written on the topic. As noted, much of the focus up to this point has been on tools and methods. Data plus mathematics and statistics gets you only machine learning, which is perfectly reasonable if that is what you are interested in, but it is not science. Again, science is about building knowledge, which requires some motivating questions about the world and hypotheses that can be brought to data and tested with appropriate methods.

Conversely, substantive expertise plus knowledge of mathematics and statistics is where most traditional research falls short. Doctoral level researchers spend most of their time acquiring expertise in a sub-field, but very little time learning about technology. This is also where much of the traditional intelligence analysis lies. Many analysts in the intelligence community have tremendous knowledge about specific parts of the world, groups and their leaders, and the culture and norms of various societies. Often, however, these analysts lack the ability to apply this knowledge to large-scale analyses using modern data-driven techniques.

Part of the reason for this is that typically the culture in academia does not value an understanding of technology. This attitude is then transferred into the Intelligence Community. Managers and intelligence veterans, however, must be open to the use of these new tools, and in doing so reward technological aspiration among analysts. Only then can the Intelligence Community move from the traditional research area to promote data science as a discipline within the analytical cycle.

Finally, I offer a cautionary word on the combination of hacking skills and substantive expertise, which I identify as the danger zone. This is where I place people who “know enough to be dangerous,” and is the most problematic area of the diagram. Those in this category may be perfectly capable of extracting and structuring data, likely related to a field they know quite a bit about. They may even have sufficient technological acumen to run a linear regression and report the coefficients; but they lack any understanding of what those coefficients mean or how to interpret them.

It is from this part of the diagram that the phrase “lies, damned lies, and statistics” emanates, because either through ignorance or malice this overlap of skills

gives people the ability to create what appear to be legitimate results without any understanding of how they got there or what they have created. The dangers of this approach are exacerbated in the context of the Intelligence Community; wherein the demand for timely analysis could supersede deliberations on the methodology applied or the interpretation of results.

Fortunately, near willful ignorance is required to have hacking skills and substantive expertise without also learning some mathematics and statistics along the way. As such, the danger zone is sparsely populated, but it does not take many to produce a lot of damage. Given the customers of intelligence products, and the stakes at play in the community, even a brief lapse into the danger zone can have catastrophic results.

As “data science,” transitions from a catchy phrase to a legitimate discipline and career path, the U.S. Intelligence Community must be able to train existing analysts in the necessary skills and to recognize new talent for recruitment. Perhaps in no other industry is the reliance on large and dynamic data streams more apparent than intelligence analysis. To maximize the value of this information, a combination of several areas of expertise are necessary, each requiring significant time and effort to master. Despite the steep curve, as the volume of data increases and the need to extract meaning becomes a permanent staple in the intelligence cycle, the community will have to work to instill the principles of data science at all levels. **Q**

Drew Conway is a Ph.D. student in the Department of Politics at New York University. Drew studies terrorism and armed conflict; using tools from mathematics and computer science to gain a deeper understanding of these phenomena. Before entering graduate school, Drew worked as an all-source analysts for the U.S. Intelligence Community where he applied his technical training to the analysis and creating specialized products.

WORKS CITED

Baker, Stephen. "Google and the Wisdom of Clouds." *Business Week*, December 13, 2007.

Harris, Mark. "Interview: Amazon's cloud computing evangelist Adam Selipsky is determined to find the silver lining in 'capital-constrained' times." *The Guardian*, March 26, 2009.

Loukides, Mike. *What is Data Science*. An O'Reilly Radar Report, O'Reilly Media, 2010.

Olson, Mike. "Hadoop: Scalable, Flexible Data Storage and Analysis." *IQT Quarterly*, Spring 2010: 14-18.