Adopting A Security Data Science Mindset

Figure 12- [f0#.eps]

Figure 12-# is a slight modification of Drew Conway’s “Data Science Venn Diagram” (<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>), a simple visualization which can help you perform a quick self-evaluation of where you currently are on your journey to becoming a security data scientist. We’ll take a look at each major component along with the interactions between some components to give you an idea of where top topics in this book fit and provide pointers for delving into or shoring up areas that may not currently be strong points for you.

The Hacker

The term “hacker” has a great deal of confusion surrounding it as it has been usurped by news media and manipulated by marketing firms. In the context of a security data scientist, “hacker” means:

* Being able to code, either scripting in a language like Python or full on programming in something like C;
* Knowing a wide variety of data formats and understanding how to slice, dice, and bend them to your will;
* Having the ability to think logically/scientifically (essentially, not jumping to conclusions) as well as algorithmically (break apart a problem into its composite parts)

If you are an information security professional who isn’t a coder, Chapters 2, 3, and 4 have been designed to help you bootstrap into that skill. If you are a coder, those same chapters cover a language that is most likely new to you (R) and place coding in the context of data analysis versus application building or systems administration, which may be more familiar problem domains for you. Whether you’re at the top of your game as a programmer or just getting started there is always more to learn and no shortage of resources available to do so, including:

**Codeacademy** (<http://www.codecademy.com/>). This is an especially good resource for those new to programming in general or those unfamiliar with a particular language. It would be worthwhile to take a look at the JavaScript & jQuery offerings given the emphasis on JSON in Chapter 8 and D3.js in Chapter 11, and if you don’t know Python well (or at all), their Python course can definitely help.

**Code School** (<https://www.codeschool.com/courses>). The offerings at Code School can be a bit overwhelming and not all are free. However, their R course is freely available at the time of writing and will help you navigate the syntax and nuances of the language.

**W3Schools** (<http://www.w3schools.com/>). If you haven’t had the opportunity to shore up your HTML/CSS/JavaScript skills, W3Schools provides an extremely friendly environment to both learn and experiment. You’ll need at least a basic understanding of these client side components if you want your analyses and results to reach the widest audience.

**StackExchange** (<http://stackexchange.com/>). While you won’t necessarily learn how to code at the StackExchange family of websites, you will have a place to look for answers or ask questions when you’re stumped. Whether it’s trying to understand some esoteric option in ggplot2 or how to do something a bit more complex with a pandas data frame there’s a very good change the answer will be in StackExchange.

When it comes to data formats, security professionals are in the unenviable position of having to be able to manipulate everything from NetFlow captures, to full packet capture (PCAP) dumps and almost every log format known to humankind. The IronPort log file snippet in the MongoDB section of Chapter 8 is an example of how “imperfect” your data world is. While that log file contains highly useful data, it’s in a format that you must parse and convert to make it useful. The only way to get good at that is to actually do it over, and over, and over again, building up reusable bits of code and techniques along the way to save you some time later on.

Learning how to think logically, scientifically and algorithmically requires time, effort and practice. Formal, in-person, instructor-led education may work best for some students, especially those who have shied away from programming. However, introductory sites like Project Euler (<http://projecteuler.net/problems>) can get you started down this path, more advanced and diverse problem sets can be found at Kaggle (<http://www.kaggle.com/competitions>), and you can delve into wide and deep security domain problems at the VAST Challenge (<http://vacommunity.org/VAST+Challenge+2013>) site (look in both current and previous years’ sections).

Overarching these three traits is the need to develop and hone a sense of *curiosity*. In fact, curiosity may be the **single most important trait** of a “hacker”. The need to know *why* or *how* something works the way it does from start to finish is an invaluable driving force when faced with a complex data science problem. When combined with the other two security data science primary skills (statistics knowledge and security domain expertise), you can eventually get to a place where developing a successful NetFlow-based malware traffic clustering algorithm is as rewarding as pwning the other team in a capture-the-flag competition.

The Statistician

Given some of the “rookie mistakes” seen in many security industry reports and the prevalence of raw counts in security dashboards there’s a high probability that statistics may be the weakest area for information security professionals. We’ve covered some statistical concepts in depth and provided a whirlwind overview of others in Chapters 4, 5, 7, and 9. A PhD in statistics is not required to be an effective security data scientist, but it’s important to have an understanding of the fundamentals of statistical analysis and machine learning, even if you’re part of a multidisciplinary team.

While you can head over to your local college or university and dive into a traditional classroom program there are two other options to consider when looking to obtain a better understanding of statistics:

Massively Open Online Courses (MOOCs) like Coursera’s Data Science course (<https://www.coursera.org/course/datasci>), edX’s Learning From Data course (<https://www.edx.org/course/caltechx/cs1156x/learning-data/1120>) and Syracuse University’s Data Science Open Online course (<http://ischool.syr.edu/future/cas/introtodatasciencemooc.aspx>) provide a low-risk way to plug into a formal statistics curriculum but aren’t right for everyone. Lectures, handouts and assignments are available at your convenience (within a course’s overall schedule) and discussion forums provide a way to interact with professors, teaching assistants, and fellow students, it can be bit overwhelming or even distracting to be in a setting with 2,000 to 4,000 individuals. Individual attention can also be difficult to obtain if you’re struggling. Employers and professional organizations may also not yet accept the certifications from MOOCs, making the time investment more for personal benefit than professional credential gains.

Online certificate or Masters courses such as UC Berkeley’s MIDS (<http://www.ischool.berkeley.edu/programs/mids>) program, University of Washington’s certificate in data science (http://www.pce.uw.edu/certificates/data-science.html) and Penn State’s Applied Statistics online curriculum (<http://www.worldcampus.psu.edu/degrees-and-certificates/applied-statistics-masters/course-list>) offer the structure and size of a traditional classroom with the convenience being online.

Understanding and applying statistics correctly is more complex than you might imagine, and individuals in disciplines with a rich history of using statistics to solve complex problems often times fall into common traps. Resources such as Alex Reinhart’s *Statistics Done Wrong* (<http://www.refsmmat.com/statistics/>) and DZone’s mis-named “Big Data” Machine Learning reference (http://refcardz.dzone.com/refcardz/machine-learning-predictive) are good to keep on hand to keep your analyses on-track.

The Security Domain Expert

When focusing on the topic of security domain expertise as it relates to data science, “thought leaders”, “gurus”, and “rock stars” need not apply. What we’re really talking about here are practitioners with solid, in the trenches, real-world experience. Depending on your area of focus (information security covers a broad range of topics), you may be applying your combined hacking skills, statistics knowledge and expertise to:

* Develop smarter endpoint protection system algorithms
* Discover new ways to detect anomalous behavior in network data
* Uncover patterns from vulnerability assessments to help determine why some systems fall out of compliance more than others
* Provide meaningful, and useful metrics for various components of your overall security program

or a host of other areas.

**Your** insight is, perhaps, **the most valuable component** to this data science triad, as it will move computations sans context into the realm of analyses driving action. There is virtually no way for an organization or individual to effectively crunch “security data” without this domain expertise. Your assistance and knowledge is vital in crafting clever questions and confirm results. Your insight into the networks and systems of your organization, behaviors and characteristics of malware, and classification and qualification incidents will be the critical factor in corresponding analyses.

The Danger Zone

A little knowledge is a dangerous thing, and having a basic ability to gather and programmatically crunch data backed by a bit of industry knowledge is where you may fall into the trap of thinking your doing data science when all you’re doing is reputational damage to all three component areas (and, potential, yourself). How do you steer clear of the danger zone?

**Embrace (versus dabble in) statistics**. Statistics and machine learning has enabled advancements in everything from a deeper understanding of the microscopic workings of our genes, to telling us how many steps and flights of stairs we’ve taken to how to build spacecraft that eventually break pas the limits of our solar system. They can absolutely help enhance our knowledge of security issues and even help solve some of them. Just don’t think you can dip your toe in. Not everyone can be a PhD in statistics, so make sure to surround yourself with a physical or virtual team with at least one strong stats person to help you stay on course.

**Dig deep but stay wide**. You will need to know certain aspects of information security just as thoroughly as individual biologists know the deep vertical segments of their discipline. But, because so many areas outside “security” (e.g. economics, politics, human rights) have an impact on “security” you’ll need to factor those in as you move from asking “What?” and “How?” to “Why?” and “Who?”. Finally, there’s a reason the CISSP certification has ten domains. You can’t be an expert in each, but you should know enough about each of them to bring in expert help when needed.

**Challenge assumptions and validate results.** Hold yourself accountable and ask others to hold you accountable all the way through your analyses. Whether you’re working on internal organizational data or performing research you intend to publish and/or speak about, pair up with practitioners who can help you keep on the straight and narrow path. When you’ve released your findings, take an example from the reproducible research (http://www.foastat.org/resources.html) movement and ensure there is sufficient documentation and data available for others to test out your findings.

Developing Developer Skills

Source Code Control

Unit Testing

Code Reviews