Security Through Data Analysis

Security with Numbers

1. Unleashing The Power Of (Security) Data

Opening chapter will describe some of the challenges in information security and why data analysis is an important component for us moving forward.

Part of this should, perhaps, draw some parallels to another discipline: medicine. Medicine moved from shamans and (for all practical purposes) faith healers to a process-based discipline to a data-based discipline, and all three of those exist simultaneously today, just like in infosec, with the data-based part being really nascent (hence the need for the book!).

There is a paper by James Shanteau from 1992 (<http://ruralgrocery.com/psych/cws/pdf/obhdp_paper91.PDF>) that covers what you are describing. He talks about role of the environment (basically how clear the feedback from an environment is) in our ability to learn from the environment and lists our several fields (including medicine) that have been transforming and improving based on improved feedback (e.g. data analysis, scientific process). The money quote there is “An insightful observation was offered by Gigerenzer (1989), who noted that historically most left-side tasks [“good” performance/feedback] began as right-side tasks [“poor” performance/feedback].” In other words, he is describing the ability for our ability to change and improve as we focus our efforts on improving methods for collecting and analyzing environmental feedback.

I’d like to make sure we acknowledge the work of those who deal with dissecting infosec “disease” data (i.e. the folks who do malware analysis) and infosec “immunity” data (i.e. the folks who find weaknesses in apps & applications) but then point out that we need the equivalent of internal CDCs and cross-organizational CDCs to both make sense of our complex data environments and help prevent malicious activity from spreading.

This would be drawing a specific analogy to medicine and I’m a little uncomfortable with that (but not too much). We should attempt to identify a picture of the “whole” data landscape and then the subcomponents and high-level data types we may come across.

We should point out areas of data analysis that have a direct impact infosec including capacity planning (e.g. “are these log servers going to be big enough for ‘n’ years growth?), malicious traffic identification, speedy forensics w/IOCs, discovering patterns (e.g. phishing campaigns, malware propogation).

Also should find a way to stress and show why it’s vital to move from being security shamans to security data scientists. -- Yup, see Shanteau above!

1. Finding Your Inner Security Data Scientist

Explain how we can do a lot of really cool analysis and tasks without needing a degree in statistics. This chapter will also outline the types of skills we will cover in this book (statistics, programming, scripting, database management and visualization techniques) and explain why each one is important and how much skill the reader should expect to develop.

BOOM: right on

1. Understanding When (Why?) 35 Is No Different Than 37

This chapter will be the first (of two) chapters on inferential statistics and will begin with a (brief, very brief) section on descriptive statistics, but jump into sample size, confidence intervals and hypothesis testing (is an observation of 35 different than another of 37?)

1. What’s The Frequency, Kenneth?

Correlation statement: how we see employee productivity in egress firewalls (correlate bandwidth usage with employee productivity), or perhaps, insider misuse with login attempts (there are often indications of “testing the water” for insiders before they go hog wild).

This chapter will cover correlation versus causation and discuss correlation techniques (pearson and scatter plots). This will lead into a high level introduction to regression techniques, but only from “what is it” discussion not necessarily how to perform and interpret regression analysis.

We can drop the R.E.M. title if you like…

1. Exploring The Dark Art Of Data Munging

This chapter will cover data sources, data collection and cleaning and/or normalizing of data. (maybe we want to talk about normalizing across disparate comparisons too here - through proportions or z-scores, etc)

While it’s a known issue to some (perhaps, many?), I think one foundational component here is talking about time, both the need for accurate/consistent timestamps and how to compensate for the lack thereof. Not accounting for that could really kill an analysis before it even begins.

It may be a good exercise to show how common log formats – like web server logs – can truly vary widely and how the intake/cleanup tasks need to account for that (perhaps also make sure to include validation steps like your VERIS python stuff). There is a very important point here to make and that is everyone, everywhere makes mistakes (the proliferation of excel problems is a good example of that). We need to point out that mistakes in data munging are not a sign of poor performance, but a natural by-product of the process and it needs to be accounted for.

It’s also probably a good place to point out that your “safe” data source may actually be malicious (i.e. log injection) and that this needs to be taken in to account when doing analysis and may also be a decision point to keep analytics “networks” and compute engines off of the main network when looking at untrusted sources.

Also consider (mebbe) pointing out that certain data bits (e.g. proxy logs) are, perhaps, sensitive and that understanding both when and how to ‘anonymize’ or mask the is necessary.

Let’s think about how to account for “security (conf/integ/avail) of your data” along with “privacy implications of data”

1. Storing and Accessing Data

This will cover traditional databases and the new breed of NoSQL solutions. Should go into strengths and weaknesses of each and help in selecting and using a data storage mechanism and hit the buzz words: Hadoop, mongo, etc.

Great place for examples and perhaps pointing out the concept of data pivoting. Traditional analysis of even a log file would shove fields into a single database for SQL processing and – depending on the purposes of analysis of the data – it may make far more sense to take portions of a data feed/log and shove some bits into one engine and some other bits into another engine, etc.

Also want to point out that we need to eat our own dog food and that access controls will be necessary considerations for various data types and point out strength/limitations of various data storage/access systems when it comes to this. [see previous chapter, we need to cover data security somewhere]

1. (Avoiding) Spatial Data

Projecting the virtual world onto the physical may not be useful. and once the pitfalls of mapping are covered, we could cover map projections and basic mapping techniques, maybe get into lat/long calculations.

Along the way here, show how to make Choropleths & pinpoint maps and what types of questions they answer/analyses they help foster.

1. Visualize This

Intro to data viz concepts, mapping data types to visualization types end up with making pretty excel/python/R charts.

This is a potentially huge chapter. I’ll ponder how to break it into more than one for a bit. Some top-of-mind stuff:

Data types that foster time-series analysis & visualization with examples in R/python/Excel

Data types that foster something like the breach data chart I made with examples

Data types that foster use of more complex chart types like spider charts

Other ones are covered in previous chapters…

Mebbe this isn’t as huge as I thought ☺

1. Making The Machine Learn For You

Intro to machine learning concepts, give 2 examples: supervised learning and unsupervised learning both from infosec. - detecting failed logins

I really like the titles of 9 & 10

<http://www.amazon.com/Machine-Learning-Mining-Computer-Security/dp/184628029X>

Unfortunately, I now have to buy this dang book (once this approach is approved)

1. Making The Machine Read For You

natural language processing - need example (not spam) from infosec here.

http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.2278

1. Back To The Future

End with a chapter on predictive analytics.