# Security Through Data Analysis

Harnessing the power of feedback through data analysis

1. Unleashing The Securing Power of Data
   1. Standing on the Shoulders of Giants
      1. We will begin by looking at other industries that have made a conversion from little-to-no data into a statistically-driven one.
   2. Shifting from Security Shaman to Data Sherpa
      1. The connections will be made between where we (in info security) are now and where the example industries were before they were transformed by data.
2. Finding Your Inner Security Data Scientist
   1. No shirt, no shoes, no degree, no problem
      1. This will explain how we can do a lot of really cool analysis and tasks without needing a degree in statistics. We will cover the “ABC” simplification method here (arithmetic, bucketing, comparing)
      2. Examples: Finding obvious anomalous firewall traffic, best IP finder service
   2. Obtaining the essential ingredients
      1. This will also outline the types of skills we will cover in this book (#1 skill is curiosity, 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 (and the rest of the chapters will go through these)
3. What’s The Frequency, Kenneth?

This chapter will be the first (of two) chapters on inferential statistics and will begin with a (brief, very brief) section on descriptive statistics. Then cover correlation versus causation and discuss correlation techniques (pearson and scatter plots). These concepts will all apply to info security and be applicable both within common tools (arcsight, other SEIM) and to more manual methods (excel).

Example Data: Proxy Logs, System logs

1. Why 35 = 37 Can Be True

This will cover a broader topic of sample size, confidence intervals and hypothesis testing (is an observation of 35 different than another of 37?) 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.

Example Data: 1) Vulnerability counts, 2) patch coverage, 3) industry reports (examples of people doing well or missing the logic)

1. Exploring The Dark Art Of Data Munging

This chapter will cover data sources, data collection and cleaning and/or normalizing of data. Special care will be given to infosec-specific data, ip addresses, domain names, timestamps, common log formats, etc.

Example Data: 1) common device log files 2) netflow data 3) windows event logs

1. Dear RDMS, It’s not you, it’s my data

This will cover traditional databases (at a high level) 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. Another important consideration here will be the security of these platforms (at a high level).

Examples: we will take the example data from Previous chapter and show the flow within various platforms (tbd).

1. It’s Spatial Data not Special 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.

Examples: 1) mapping bots, 2) challenges in geo-location from IP/attackers 3) ASN.1 pitfalls and gotchas

1. Let’s get visual, visual

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

We will take previous examples (as the readers should be familiar with the data) and create visualizations from those.

1. Making The Machine Learn For You

Intro to machine learning concepts, give 2 examples: supervised learning and unsupervised learning both from infosec and cover some areas of machine learning like naïve bayes, etc (while we will discuss spam filtering will probably avoid the spam example as it’s overused and covered elsewhere).

Examples: detecting failed logins, malware detection and classification (naïve bayes filter)

1. Making The Machine Read For You

Intro to natural language processing - need example (not spam) from infosec here. NLP has many challenges and benefits, will touch upon these at a high level.

Example: unstructured data classification for Data loss prevention (DLP)

1. Prediction is hard, especially about the future

End with a chapter on predictive analytics. Discuss

Capacity planning: modeling growth in centralized logging services, Predicting rogue behavior (insider misuse) from a small set of features (there has been some interesting research here on detecting at-risk employees).

1. Keeping it Simple

Discuss what we’ve talked about and a few final areas for areas we could study more, but the real point of this chapter is to bring all of these points back to the reality of our environments.