Security Through Data Analysis : Harnessing The Power Of Feedback

Introduction

Chapter 1: Unleashing The Securing Power Of Data

Standing On The Shoulders Of Giants

We will begin by looking at other industries that have made a conversion from little-to-no data into a statistically driven one.

**Use Case**: Agriculture -> Agri-infomatics

**Use Case:** Biology -> Bioinformatics/Epidemiology

Shifting From Security Shaman to Data Sherpa

The connections will be made between where we (information security professionals) are now and where the above example industries were before they were transformed by data.

Chapter 2: Finding Your Inner Security Data Scientist

No Shirt; No Shoes; No Degree? No Problem!

This section will explain how we can do a significant amount of extremely useful analysis and tasks without needing a degree in statistics. We will introduce the “ABC” (Arithmetic, Bucketing & Comparing) Simplification Method here.

**Use Case:** Discovering Anomalous Firewall Traffic

**Use Case:** Identifying The Cost Of Two-Factor Authentication

Understanding The Essential Skills & Ingredients

This will outline the types of skills: curiosity, statistics, programming, scripting, database management and visualization techniques and will also emphasize that security domain knowledge is one of the most essential components in successful data analysis.

Chapter 3: Learning The “Hello World” Of Security Data Analysis

Deciphering The Not-so-secret Secret Of Data Analysis: Data Munging

This section will delve into the fundamentals of acquiring security data and how to manipulate it so that it can be analyzed. A major focus will be teaching security analysts how to setup a repeatable data analysis toolkit and workflow.

**Use Case:** Swimming Unified Log Stream (CLF)

Exploring The Dark Art Of Data Munging

This section will dig deeper into the topic and cover topics involving date/time stamps, dealing with data from multiple sources and determining when you need more than just “security data” to solve a problem.

**Use Case**: Normalizing NetFlow Data

**Use Case**: Analyzing Windows Event Logs

**Use Case**: Helping The Help Desk

Chapter 4: Tuning The Right Frequency: Security Analysis By The Numbers

Learning From Our Frequentist Forefathers

This section will detail how to explore data with descriptive statistics and significantly expand on the “ABC” methods introducted in Chapter 2.

Separating Correlation From Causation

Much like the sister disciplines covered in Chapter 1, security can fall prey to the pitfalls of assigning causation when there is just correlation. Correlation is, however, a key component in security data analysis and this section will walk through use cases that will help avoid the pitfalls and tie the right elements together.

**Use Case**: Productivity In Proxy Logs

**Use Case**: Whitehat Statistics Report

**Use Case**: Security Event Correlation

Chapter 5: Knowing When 35 == 37: Finding The True Messages In Security Data

Measuring The “Power” Of Sample Size

This section will cover more statistical methods as the relate to understanding what the security data is saying (and how strongly or weak those messages are).

**Use Case**: Vulnerability Counts

**Use Case**: Security Patch Coverage

Being Secure In Your Uncertainty

Security data and analysis—much like most other disciplines—is imperfect and understanding how to compute and convey uncertainty is a key element very often missed or just left out of our discipline. This section will walk through examples to show how to identify, embrace and communicate it.

**Use Case**: Trustwave’s Industry Report Trending Year-over-year

**Use Case**: Malware Gone Wild (Using inferential statistics to detect a malware outbreak *before* it gets crazy)

Chapter 6: Breaking Up With Your Relational Database

Realizing The Container Has Constraints

Security analysts tend to have a very monolithic view of data and the “giant SQL database” view of the world actually really hampers analysis and understanding. This section will show how to make the most of traditional data stores for security analysis.

**Use Case**: MySQL Memory (And Other) Tables

Managing Non-relational Data (Saying “Yes” To NoSQL)

This section will introduce NoSQL systems (e.g. Hadoop/PacketPig, MongoDB, Couch, Redis) and explain conditions when they can/should be used. This section will also cover the concept of data as a “stream” that can be tapped into in different ways to perform different analyses.

**Use Case**: Storing And Accessing Netflow Data (Continuation of Chapter 3 Analysis)

Chapter 7: Visualizing Security Data

Building The Foundation Of Security Data Visual Analysis And Communication

This section will briefly cover the foundational visual communication methods as they relate to analyzing and communicating security data. It will cover different tools from Excel, to Python, to R

**Use Case**: Graphing Netflow Data With [Excel|Python|R] (expansion of Chapter’s 3 & 6 analysis)

**Use Case**: Improving Visual Defaults in [Excel|Python|R]

**Use Case**: Visualizing System Log Data (expansion from Chapter 3)

Chapter 8: Mapping Security Data

Realizing That Spatial Data Is Not Necessary Special Data

This section will cover the fundamentals of performing Geo-IP mapping (which is a fundamental component of many areas of security analysis) and also explain the pitfalls of Geo-IP mapping.

**Use Case**: The Many Challenges In Geo-location

**Use Case**: Mapping Botnets

Mapping Outside The Continents

Network data does not necessarily have to be conveyed on a map and this section will walk through a use case that shows how to discover patterns and clusters with mapping tools that aren’t geographically centered.

**Use Case**: Mapping Malicious Clusters With ASN.1 Data

Chapter 9: Making The Machine Learn For You

De-mystifying Machine Learning

This section will explain what ML is and is not and provide the necessary background for the next section.

Understanding The Security Potential Of Machine Learning

This section will walk through use cases that will help the reader truly understand how to apply machine-learning concepts to security.

**Use Case**: Discovering Account Takeovers (An example of “supervised learning”)

**Use Case**: Detecting And Classifying Malware With Naïeve Bayes Networks

Chapter 10: Making The Machine Read For You

Introduction To Textual Analysis (NLP)

This section will cover the basics of text analysis and provide the background necessary for the next section.

Using NLP In Security

There is a surprisingly large amount of security relevant content in web pages, Word documents, e-mail systems, etc. This section will walk through use cases covering data loss detection/prevention and using text analysis concepts with other traditional statistical processing to perform attribution of anonymous blog/forum posts.

**Use Case**: Using NLP in DLP (Data Loss Prevention)

**Use Case**: Who Wrote That? (Attribution of anonymous blog/forum posts)

Chapter 11: Predicting The Future With Security Data

Can You Do Better Than Punxsytawney Phil?

This section will relate that everyone is jumping into the prediction game and most folks are doing it very poorly. While predictive analytics is a full-on book in-and-of itself, this chapter is designed to walk through use cases to show the basics of how to do predictive analytics and also to show when you can and should (and can’t and shouldn’t) do PA.

**Use Case**: Modeling growth in centralized security logging systems

**Use Case**: Predicting rogue behavior (insider misuse)

Chapter 12: Keeping It Simple

Putting Security Data Analysis Into Perspective

This section will take a step back from the “drilling for oil” approaches in the previous chapters and introduce a “mile wide/inch deep” approach to analyzing security data (i.e. you can pan for gold in a stream sometimes vs drill for oil).

Understanding The Reality Of Our Environments

A final call to understand that our data isn’t perfect (and what that means for analysis & visualization), our methods are nascent and improving and that our discipline has much to learn from other that have gone before and that are continuing to evolve.