Security Through Data Analysis : Harnessing The Power Of Feedback

Introduction

Chapter 1: Unleashing The Securing Power Of Data

Standing on the shoulders of giants

Looking at other industries that have made a conversion from little-to-no data to being data-driven.

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

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

Skills of a Security Data Scientist

Domain (security) knowledge, programming, data manipulation, statistics, visualization

Choosing a good research question

Exploratory Data Analysis

**Use Case:** Zero Access (negative finding, still value and limits of EDA)

Can show comparison of income levels to bot infections, cuz it’s funny.

Chapter 2: Learning The “Hello World” Of Security Data Analysis (2 chapters?)

(heavy focus on “risk” in this data: it sets up veris)

Describing an IP address

Descriptive visualization of just IP addresses (as in map of internet)

**Use Case:** AlienVault list of badness: geo-location, ASN

Can we get lists over time (for comparisons and shifts)

Using Python to lookup geolocation and/or ASN, R to graph it

Answering the “So what” (using infosec to process it)

(( DNS records able to be tied in here? ))

Merging data for more value

Using more than just “security data” to solve security problems

Scraping the web for fun and data

**Use Case:** Pulling data from ASN providers to compare to “badness” (catalog of denominators)

(was ch 8) Realizing that spatial data may not be special data

Performing Geo-IP mapping (and the gotchas in doing so)

**Use Case**: Generic geo-location of IP addresses

**Use Case**: Mapping Botnets

Mapping outside the continents

Discovering patterns and clusters with mapping tools

**Use Case**: Mapping malicious ASN.1 data

Chapter 3: Descriptive Visualization

Describing attributes of IP addresses over time

Use case: Opportunistic Network Traffic (tcpdump)

Time series analysis

Descriptive Statistics (average packets per day over time)

5-number summaries, and other descriptive stats

Describing Visually

Scatter plots of traffic

Creating a video of packets

Chapter 4: Breaking Up With Your Relational Database

Realizing the container has constraints

**Use Case**: MySQL memory (and other) tables

Managing non-relational data (saying “Yes” to NoSQL)

Explaining alternative data stores and their strengths:

Hadoop/PacketPig, MongoDB, Couch, Redis, etc

**Use Case**: Storing and accessing netflow data (continuing ch.3 data)

**Perhaps**: use firewall, IDS or other simple data

Chapter 5: It’s a messy, messy, messy world

Applying the same descriptive visualization method to other data

Introduction to VERIS

Who does what to who and what and how (and so what)

**Use Case**: VERIS Community Database (vcdb)

The challenges of information sharing

The cost-per-datum challenge

**Use Case:** The misaligned research question

Inferential estimations

Confidence intervals, fisher-test and chi-square for independence

Comparisons to other methods

Strengths and Limitations (we all got me)

Chapter 6: Visualizing Your Security Data

Building the foundation of security data visual analysis and communication

How-to examples each in in Excel, Python and R

**Use Case**: Graphing trends in netflow data (expansion of ch 3 & 6 analysis)

**Use Case**: Improving visual defaults

**Use Case**: Visualizing system logs (expansion from ch 3)

Chapter 7: Having the machine learn for you

De-mystifying machine learning

Basics of ML with the necessary background for the next section

Understanding the (security) potential of machine learning

Unsupervised Learning: Clustering host activity

Techniques: MDS, Knn, k-means

**Use Case:** Proxy/FW traffic, netflow?

Supervised Learned: Classifying host activity

Techniques: Logistic regression, random forest

Chapter ??: Making The Machine Read For You

Introduction to textual analysis (NLP)

Using NLP in security

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

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

Chapter 8: Bringing it all together in a Dashboard

How a dashboard is not a simple visualization

All the dashboard stuff here

Chapter 9: Keeping It Simple

Putting security data analysis into perspective

Comparing a “drilling for oil” approach to a “pan for gold” approach

Understanding the reality of our environments

Reiterating that data analysis assists our thinking, it does not replace it