Chapter 1: Unleashing The Securing Power Of Data (Jay)

*[Setting the stage: Use cases showing real people and real problems being solved by data (outside infosec), discuss skills of a security data scientist and ways to approach data analysis]*

Standing On The Shoulders Of Giants

Snow, Fisher and Demming

Developing the Skills Of A Security Data Scientist

Curiosity

Domain (security) knowledge

Programming

Data management

Statistics (in the broad sense)

Visualization/Communication

Recognizing The Importance of a Good Question

Choosing A Good Research Question

Exploratory Security Data Analysis: Seeking questions we can answer

Looking ahead: How to read this book

WEB CONTENT: visualizations without code

Chapter 1.5: Building Your Analytics Toolbox: A Primer on Using R & Python for Security Analysis (Bob/Jay)

*[Enabling the reader to set up their own environment locally to follow along with our examples and data sets]*

About This Chapter

Why Python; Why R; ***And***, Why Both?

Setting Up Your R Environment

Finding and Installing R and RStudio

Installing Packages with CRAN and devtools

Knowing when **not** to upgrade

Getting Familiar With R Syntax and Commands

Jumpstart Your Python Analytics With Anaconda

Installing Packages with pip and easy\_install

Knowing when **not** to upgrade

Understanding iPython and the SciPy Ecosystem

Getting Familiar with Python Syntax

Where To Go To Learn More

WEB CONTENT: full set of links to R & Python resources

Chapter 2: Learning The "Hello World" Of Security Data Analysis (Bob)

*[Our first data set, we will cover the basics of handling data, reading it in python and R, and doing some basic exploration and descriptive statistics]*

Chapter Use Case: Parsing AlienVault IP Reputation Data

Reading in data

Getting a feel for the data

Basic Stats

Basic Graphs

Web scraping

Augmenting your analysis with different, relevant data

Asking a basic question of the data

Which country has the most malicious hosts of type “x”

WEB CONTENT: AlienVault data (w/link to source), python code, R code, visuals

Chapter 3: Analyzing “Badness” (Bob)

*[Same data as previous chapter, but now we’ll really dig in and pull out some interesting relationships in the data and create a few visualizations]*

Chapter Use Case: Crunching AlienVault IP Reputation Data

Dissecting the “IP Address”

32-bit integer (“how does your computer see an IP address?”) + machine info

Part of a subnet / logical layout / MAC addresses, perhaps has a hostname (DNS)

Larger context: part of a global network organized by ASNs (BGP)

Lager context: Has a physical location

Mapping Outside The Continents

USE CASE: Visualizing AlienVault ASN data (force-directed network graphs of malhost ASN groupings)

WEB CONTENT: AlienVault data (w/link to source), python code, R code, visuals

Chapter 4: Mapping “Badness” (Jay)

*[Covering spatial data, what can we do with geolocation of IP addresses? Aside from learning how to create and manipulate maps, main lesson is that geo-location isn’t all that interesting unless we correlate other factors]*

Chapter Use Case: ZeroAccess Botnet Analysis (note: working on another data set)

What can you learn from just a set of lat/lon pairs?

Getting basic lat/long metadata (lat/lon -> country/city)

Visualizing lat/long data

Choropleth

Dot plot

Getting more advanced metadata (internet user population & income)

The quest for correlation

WEB CONTENT: ZeroAccess code & visuals

Chapter 5: Improving Your Security-oriented Visualizations (Jay)

*[This introduces the core concepts of data visualization using firewall session data for the examples; this will set up future discussions on both visualizations and the data set]*

Chapter Use Case: Exploring Your Firewall Data (Jay: Severski’s Data)

Understanding The Foundations Of Good Visual Communication

Position, Length, Angle, Slope, Area, Volume, Color Saturation and Hue

Addressing the silent minority (colorblind)

Leverage use case to run through basics and then comparisons of core (bar/dot/line/pie/scatterplots) charting techniques

Remembering the communication

Leveraging more than visuals, tables and words

Visualizing tabular data

WEB CONTENT: [visual defaults use case] R code, python code

WEB CONTENT: Firewall data set, R code, visuals

Chapter 6: Getting A Handle On Your Security Data With Descriptive Statistics And Descriptive Visualization (Jay)

*[Advanced visuals with time-series data and distributions; visuals as a descriptive tool for univariate (or even multivariate) data. Will also create a video or two from the firewall data]*

Chapter Use Case: Expanding from Chapter 5 with Severski’s Data

Describing Data Over Time

Moving averages

Time-series plots

Probability density function

Describing Collections of Data with Numbers

How to use descriptive statistics in a security context

distributions, probability density, quantiles, measures of center, standard deviation and skew

Describing Collections of Data Visually

How to use descriptive visualizations in a security context

Box plots, histograms, etc

Moving with Animations

Understanding The Challenges Of Visualizing Lots Of Data

Radial Graphs Example and other hair-ball diagrams

WEB CONTENT: R code, visuals

Chapter 7: Learning From Security Breaches (Bob)

*[This chapter will talk about collecting data from processes rather than logs and how we can learn from failures by collecting and analyzing breach data]*

Turning Chaos Into Structure

The Power Of Structured Recording During An Incident

Understanding & Using VERIS

Comparisons To Other Methods (Strengths and limitations)

Being Cautious About Inferential Estimations

USE CASE: Visualizing VERIS Community Database (Breach Data)

Callout: The Cost-per-datum Challenge

Looking At And Learning From Other Community Breach Data Sets

USE CASE: PRC Aggregated Breaches

WEB CONTENT: breach data, R code, visuals

Chapter 8: Breaking Up With Your Relational Database (Bob)

*[We haven’t touched on data management too much, but this chapter will dive into different approaches to storing and accessing data and the options we have and the recent advancements that open up our options]*

Realizing The Container Has Constraints

Understanding The Limitations Of A Monolithic Data Store

Tables can introduce unnecessary complexity

Exploring Alternative Data Stores

Survey of core SQL alternatives, identifying strengths and uses each in context

USE CASE: "Have we seen this IP address?"

practical example of how a traditional monolithic approach can hinder use of critical threat intelligence and how re-thinking how you intake, crunch and store data can open up new possibilities

WEB CONTENT: sample code for the use case

Chapter 9: Having The Machine Learn For You (Jay)

*[This will give an introduction to Machine Learning in the context of either netflow data or the same firewall data from chapters 6 and 7. Purpose here is to show methods of ML, not to do fancy analysis here]*

De-mystifying Machine Learning

Will discuss the surprisingly straightforward underpinnings of ML and setup the rest of the chapter

Understanding The Security Potential of ML

Unsupervised Learning: Clustering Host Activity

Having the computer figure out the patterns we can’t see

Applying Multi-dimensional scaling and visualizing Euclidian distances

K-nearest neighbors method

K-means method

Learning point: machine can profile “good” and “bad” behaviors

Supervised Learning: Classifying Host Activity *(note: may not do these, depending on data sources)*

Training the computer on patterns we see

Logistic regression

Random forests

Chapter 10: Building Dynamic Security Dashboards (Bob)

*[This will bring together many of the concepts we’ve covered, from some basic descriptive visualizations to distributions and machine learning. Will also introduce a few core concepts specific to dashboards]*

Designing Dashboards For Effective Security Response

Dashboards are a call to action

Making differences stand out

The never-ending quest for "so what?"

Applying Appropriate Visualizations To Your Security Data Streams

Knowing when and how to use line graphs, bar charts, maps, etc

Understanding The Importance of Baselines And Thresholds

Communicating With Dashboards

There is no One Security Dashboard to rule them all; designing security dashboards

USE CASE: The Incident Response Manager's Dashboard

At-a-glance overview

Current incidents in play and broad view of incidents over time

Insight about the incident handler team members

USE CASE: The Threat Management Dashboard

Tactical overview of internal threat landscape

Communication attempts (success/fail) to "badness" (i.e. Matched IoCs)

to IPs/ASNs

from what (servers/workstations)

possibly geo-located

Successful malware infections trends w/emphasis on known 0-days

Chapter 11: Keeping It Simple (Bob/Jay)

Putting Security Data Analysis Into Perspective

Comparing A "Drilling For Oil" Approach To a "Pan For Gold" Approach

Understanding The Reality Of Our Environments

Re-iterating That Data Analysis Assists Our Thinking, Not Replaces It

What Lies Ahead In Security Analytics?