Measurable Security: Information Security Using Data Analysis, Visualization, and Dashboards

Introduction

How we learned to love data

Acknowledgements

Chapter 1: Unleashing The Securing Power Of Data

Standing on the shoulders of giants

I think we’d want a few anecdotes here to demonstrate the power of learning from data. (rather important out of the gate)

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

#### Really important to establish why people will want to do this, or bad things if they don’t.

#### Is data science expensive?

A brief history of risk analysis/data analysis

#### Calculations are enabled with the introduction of the Arab-Hindu numbering system, though introduced by Fibonacci in 13th century, it wasn’t until the printing press of the 15th century it was adopted.

#### gambling drives this notion of “risk” (Pascal)

#### Pascal and Fermat stumble into probability (which is still difficult to understand today)

#### Pearson wants to “count all the things” thinking that knowledge can be gained if we just get enough things to measure

#### Fisher busts the chops of Pearson, and says the key is random sampling

#### There is the birth of the computer (which is essential in this story)

#### Tukey introduces [exploratory] data analysis and visualization as an alternative to classical statistics, early pioneer of computers and visualization.

#### Cleveland coins “data science”, split of Machine Learning and Data Mining.

Shifting from security Shaman to data Sherpa

Linking information security with other pre-data-driven industries

Data analysis assists our thinking, it does not replace it

Data analysis assists our thinking, it does not replace it

Chapter 2: Finding Your Inner Security Data Scientist (Jay)

No shirt, no shoes, no degree? No problem!

Performing basic analysis without advanced statistics using the “ABC” (Arithmetic, Bucketing & Comparing) simplification method

…and creating graphics from these

**Use Case:** Discovering anomalous firewall traffic

**Use Case:** Identifying the cost of two-factor authentication

Obtaining the essential skills & ingredients

Discussion of skills: curiosity, statistics, programming, scripting, database management and visualization techniques

Getting real benefit from combining with domain knowledge in information security

Using Descriptive Statistics

#### Central Tendancy: Mean, mode, median

#### Variance - (this is something we really want to focus on, as I wrote in sample chapter, I think computer engineers see variance as “imperfect” or “unpredictable” and that’s something we really want to dispel.

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

Deciphering the not-so-secret secret of data analysis: Data munging

How to acquire and prepare security data

Creating a repeatable data analysis toolkit and workflow

**Use Case:** Swimming the river of log data (CLF)

Exploring the dark art of data munging

The importance of timestamps and handling time series data

Combining multiple data sources

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

**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 (Jay)

Learning from our frequentist forefathers

Shifting from descriptive to inferential statistics and outgrowing the ABC method (ch. 2)

Separating correlation from causation

Seeing attack patterns that do not exist

**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 (Jay)

Measuring the “power” of sample size

Strong versus weak messages in the data

**Use Case**: Vulnerability counts

**Use Case**: Security patch coverage

Being secure in your uncertainty

Embracing hypothesis testing and confidence intervals with open eyes

**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

**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)

Chapter 7: 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)

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 8: Dashboards

Realizing

Performing

**Use Case**:

**Use Case**:

Mapping

Discovering

**Use Case**:

Chapter 9: Making 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

**Use Case**: Discovering account takeovers (with “supervised learning”)

**Use Case**: Detecting and classifying malware with Naïeve Bayes networks

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 10: Data Mining

Chapter 11: Predicting The Future With Security Data

Can you do better than a groundhog?

The hype and hopes of predictive analytics

How to perform basic predictive analysis

**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

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

Appendix: List of Data Prep/Cleaning tools