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

**Introduction**

How we learned to love data

(Jay note: there is a relation between data analysis and what we talk about as security “metrics” and/or “risk”, wondering where we talk about that, intro perhaps?)

Acknowledgements

1. **Chapter 1: Unleashing The Securing Power Of Data**
   1. Standing on the shoulders of giants
   2. **Use Case**: Agriculture -> Agri-infomatics
      1. <http://www.nass.usda.gov/About_NASS/History_of_Ag_Statistics/>
      2. Fisher taking over Rothamsted and 90 years worth of “useless” data. The data would record amount of “artificial manures” (fertilizers) and weather for that growing season. They were aggregated using “fertility indexes” (weighted scoring I think) for combining the data points. Fisher came in and showed that the impact of weather was much greater than anyone was accounting for and that most of the data was useless and misleading.
      3. <http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15171/1/16.pdf>
      4. - Wish I could find more discussion of this period in fisher’s life.
      5. What was the “determinism” in 19th century science (and mathematical models)?
   3. **Use Case:** Biology -> Bioinformatics/Epidemiology
      1. John Snow
   4. **Notes**: I think we’d want a few anecdotes here to demonstrate the power of learning from data. (rather important out of the gate)
      1. Looking at other industries that have made a conversion from little-to-no data to being data-driven.
      2. Really important to establish why people will want to do this, or bad things if they don’t.
      3. Is data science expensive?
   5. **A brief history of risk analysis/data analysis**
      1. 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.
      2. gambling drives this notion of “risk” (Pascal)
      3. Pascal and Fermat stumble into probability (which is still difficult to understand today)
      4. Pearson wants to “count all the things” thinking that knowledge can be gained if we just get enough things to measure
      5. Fisher busts the chops of Pearson, and says the key is random sampling
      6. There is the birth of the computer (which is essential in this story)
      7. Tukey introduces [exploratory] data analysis and visualization as an alternative to classical statistics, early pioneer of computers and visualization.
      8. Cleveland coins “data science”, split of Machine Learning and Data Mining.
   6. Shifting from security Shaman to data Sherpa
   7. Linking information security with other pre-data-driven industries
   8. Data analysis assists our thinking, it does not replace it
   9. Data analysis assists our thinking, it does not replace it
2. **Chapter 2: Finding Your Inner Security Data Scientist (Jay)**
   1. *No shirt, no shoes, no degree? No problem! (no training necessary)*
      1. Discussion of skills: curiosity, statistics, programming, scripting, database management and visualization techniques
      2. Counting is not the first step, finding the right question is the first step (this is decision support)
         1. Getting real benefit from combining with domain knowledge in information security
   2. **Use Case:** Discovering anomalous firewall traffic
      1. Alternatives: brute force an AES key, brute force a pw
   3. **Use Case:** Identifying the cost of two-factor authentication
   4. **Beyond just counting:** summarizing data (descriptive stats)
      1. Central Tendancy: Mean, mode, median
      2. 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.
      3. Other: Skew/kurtosis, std dev (and more, just mention in passing)
      4. Mention Pearson again with his “count all the things” approach
      5. Mention Wheelen (p6), “an overreliance on any descriptive statistic can lead to misleading conclusions, or cause undesirable behavior” -- we cannot drive our decision making on simplifications
      6. Introduce the transition into inferential statistics
3. **Chapter 3: Learning The “Hello World” Of Security Data Analysis (Bob)**
   1. Deciphering the not-so-secret secret of data analysis: Data munging
      1. 90% of the time of effort in data analysis is not spent on analyzing data.
   2. Creating a repeatable data analysis toolkit and workflow
   3. **Use Case:** Swimming the river of log data (CLF)
      1. The importance of timestamps and handling time series data
      2. Using more than just “security data” to solve security problems
         1. Combining multiple data sources
         2. Do this in the context of the use case
   4. **Use Case**: Normalizing NetFlow Data
   5. **Use Case**: Analyzing Windows Event Logs
   6. **Use Case**: Helping The Help Desk
4. **Chapter 4: Tuning The Right Frequency: Security Analysis By The Numbers (Jay)**
   1. **Note: this chapter will be about “classic statistics”**
   2. **Inferential statistics**
      1. Sample is to population as statistics is to parameter
   3. **Central Limit Theorem (can we build an example here?)**
      1. Not sure how deep we want to go
   4. **Correlation/Causation, Sample Bias and Sample Error**
      1. Error is not a mistake, bias is not a matter of opinion
      2. Discuss limitations of inference (Causation comes from outside the data)
      3. discuss data collection pitfalls and limitations
         1. we will always have error and we can account for it
   5. **Inference about a single variable**
   6. **Use Case**: Productivity in proxy logs
   7. **Use Case**: Whitehat statistics report
   8. **Use Case**: Security event correlation
   9. **Introduction to Hypothesis testing**
5. **Chapter 5: Knowing When 35 == 37: Don’t be deceived by numbers (Jay)**
   1. Probability (A very important topic)
      1. “Why do so many people find [probability](http://www.wired.co.uk/news/archive/2010-08/18/probabilistic-chip) theory so unintuitive and difficult? After years of careful study, I have finally found it's because probability is unintuitive and difficult.” http://www.wired.co.uk/magazine/archive/2011/09/ideas-bank/david-spiegelhalter-probability-is-likely-to-confuse-people
      2. Law of Large Numbers
      3. Probability within Hypothesis testing
      4. Taleb notes that fisher was right: “Fisher proposed *p* as an informal measure of evidence against the null hypothesis. He called on researchers to combine *p* in the mind with other types of evidence for and against that hypothesis”
      5. base-rate?
      6. Bayesian?
      7. birthday paradox? (probability of guessing crypto key)
      8. binomial probability
      9. “perfect secrecy” within cryptography is matching probabilities.
   2. Measuring the “power” of sample size
   3. Strong versus weak messages in the data
   4. **Use Case**: Vulnerability counts
   5. **Use Case**: Security patch coverage
   6. Being secure in your uncertainty
   7. Embracing hypothesis testing and confidence intervals with open eyes
   8. **Use Case**: Trustwave’s industry report: trending year-over-year
   9. **Use Case**: Malware gone wild (using inferential statistics to detect a malware outbreak *before* it gets crazy)
6. **Chapter 6: Breaking Up With Your Relational Database**
   1. Realizing the container has constraints
   2. **Use Case**: MySQL memory (and other) tables
   3. Managing non-relational data (saying “Yes” to NoSQL)
   4. Explaining alternative data stores and their strengths:
   5. Hadoop/PacketPig, MongoDB, Couch, Redis, etc
   6. **Use Case**: Storing and accessing netflow data (continuing ch.3 data)
7. **Chapter 7: Visualizing Your Security Data**
   1. Building the foundation of security data visual analysis and communication
   2. How-to examples each in in Excel, Python and R
   3. **Use Case**: Graphing trends in netflow data (expansion of ch 3 & 6 analysis)
   4. **Use Case**: Improving visual defaults
   5. **Use Case**: Visualizing system logs (expansion from ch 3)
   6. Realizing that spatial data may not be special data
   7. Performing Geo-IP mapping (and the gotchas in doing so)
   8. **Use Case**: Generic geo-location of IP addresses
   9. **Use Case**: Mapping Botnets
   10. Mapping outside the continents
   11. Discovering patterns and clusters with mapping tools
   12. **Use Case**: Mapping malicious ASN.1 data
8. **Chapter 8: Dashboards**
   1. Realizing
   2. Performing
   3. Use Case:
   4. Use Case:
   5. Mapping
   6. Discovering
   7. Use Case:
9. **Chapter 9: Making The Machine Learn For You**
   1. De-mystifying machine learning
   2. Basics of ML with the necessary background for the next section
   3. Understanding the (security) potential of machine learning
   4. **Use Case**: Discovering account takeovers (with “supervised learning”)
   5. **Use Case**: Detecting and classifying malware with Naïeve Bayes networks
   6. Introduction to textual analysis (NLP)
   7. Using NLP in security
   8. **Use Case**: Using NLP in DLP (Data Loss Prevention)
   9. **Use Case**: Who Wrote That? (Attribution of anonymous blog/forum posts)
10. **Chapter 10: Data Mining**
11. **Chapter 11: Predicting The Future With Security Data**
    1. Can you do better than a groundhog?
    2. Simple prediction: monte carlo simulation
    3. The hype and hopes of predictive analytics
    4. How to perform basic predictive analysis
    5. **Use Case**: Modeling growth in centralized security logging systems
    6. **Use Case**: Predicting rogue behavior (insider misuse)
12. **Chapter 12: Keeping It Simple**
    1. Putting security data analysis into perspective
    2. Comparing a “drilling for oil” approach to a “pan for gold” approach
    3. Understanding the reality of our environments
    4. Reiterating that data analysis assists our thinking, it does not replace it