Introduction

"It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to."

Bilbo Baggins, The Fellowship of the Ring

italic

In recent years, cybersecurity has taken center stage in the personal and professional lives of the majority of the global population. Data breaches are an almost a daily occurrence, and intelligent adversaries target consumers, corporations, and governments with practically no fear of being detected or facing consequences for their actions. This is all occurring while the systems, networks, and applications that comprise the backbones of commerce and critical infrastructure are growing ever more complex, interconnected, and unwieldy.

delete: an

Delete: almost

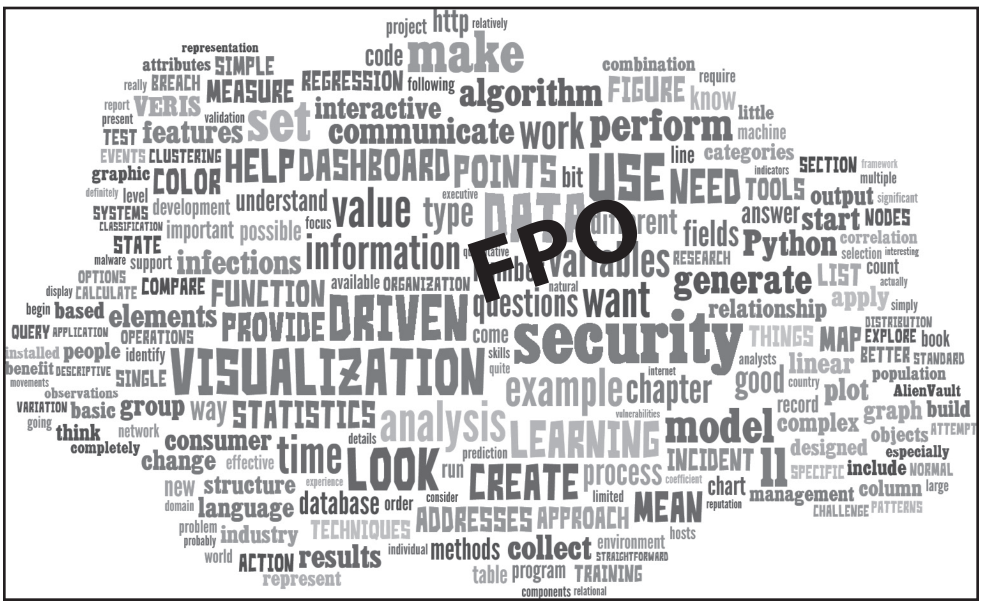
Defenses built solely on the elements of faith-based security—unaided intuition and "best" practices—are no longer sufficient to protect us. The era of the security shaman is rapidly fading, and it's time to adopt the proven tools and techniques being used in other disciplines to take an evolutionary step into *Data Driven Security*.

hyphen

Overview of the Book and Technologies

Data Driven Security: Analysis, Visualization and Dashboards has been designed to take you on a journey into the world of security data science. The start of the journey looks a bit like the word cloud shown in Figure 1, which was created from the text in the chapters of this book. You have a great deal of information available to you, and *may* be able to pick out a signal or two within the somewhat hazy noise on your own. However, it's like looking for a needle in a haystack without a magnet.

hyphen



Au: delete "FPO" in Figures 1 & 2

Au: Book has figure captions. Captions for Figs. 1, 2 wanted here?

Figure 2

s/b I-1

You'll have much more success identifying what matters (see Figure 2) if you apply the right tools in the most appropriate way possible.

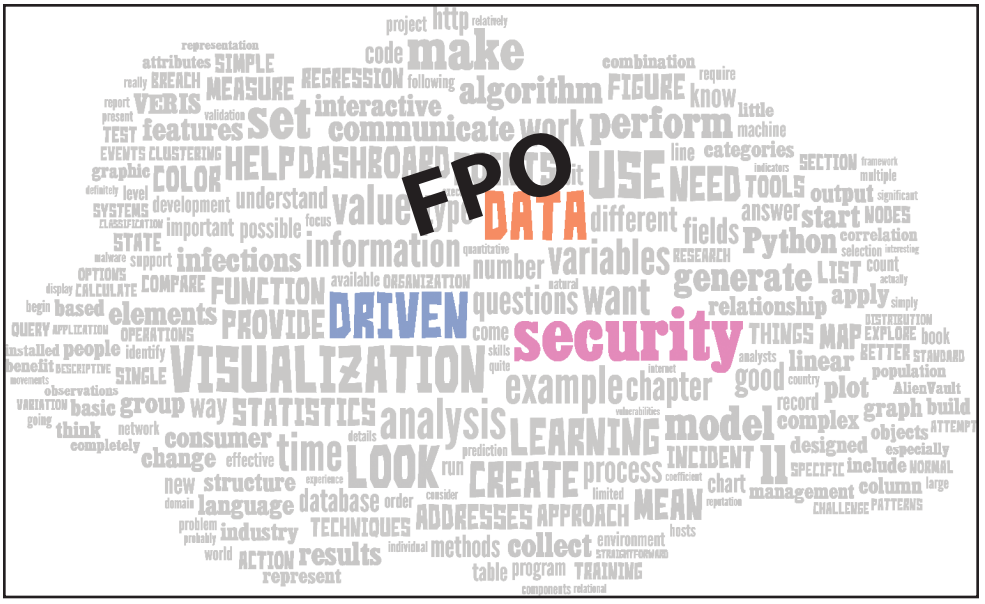


FIGURE 2

This book focuses on Python and R as the foundational data analysis tools, but also introduces the design and creation of modern static and interactive visualizations with HTML5, CSS, and JavaScript. It also provides background on and security use cases for modern NoSQL databases.

How This Book Is Organized

Rather than have you gorge at an all-you-can-eat buffet, the chapters are more like tapas—each with their own distinct flavor profiles and textures. Like the word *tapas* itself suggests, each chapter covers a different foundational topic within security data science and provides plenty of pointers for further study.

Chapter 1 lays the foundation for the journey and provides examples of how other disciplines have evolved into data-driven practices. It also provides an overview of the skills a security data scientist needs.

Chapters 2, 3, and 4 dive right into the tools, technologies, and basic techniques that should be part of every security data scientists' toolbox. You'll work with AlienVault's IP Reputation database (one of the most thorough sources of malicious nodes publicly available) and take a macro look at the ZeuS and ZeroAccess botnets. We introduce the analytical side of Python in Chapters 2 and 3. Then we thrust you into the world of statistical analysis software with a *major* focus on the R language in the remainder of the book. Unlike traditional introductory texts in R (or statistics in general), we use security data throughout the book to help make the concepts as real and practical as possible for the information security professional.

Chapter 5 introduces some techniques for creating maps and introduces some core statistical concepts, along with a lesson or two about extraterrestrial visitors.

Chapter 6 delves into the biological and cognitive science foundations of visual communication (data visualization) and even shows you how to animate your security data.

This lays a foundation for learning how to analyze and visualize security breaches in **Chapter 7**, where you'll also have an opportunity to work with real incident data.

Chapter 8 covers modern database concepts with new tricks for traditional database deployments and new tools with a range of NoSQL solutions discussed. You'll also get tips on how to answer the question, "Have we seen this IP address on our network?"

Chapter 9 introduces you to the exciting and relatively new world of machine learning. You'll learn about the core concepts and explore a handful of machine-learning techniques and develop a new appreciation for how algorithms can pick up patterns that your intuition might never recognize.

Chapters 10 and 11 give you practical advice and techniques for building effective visualizations that will both communicate and (hopefully) impress your consumers. You'll use everything from Microsoft Excel to state of the art tools and libraries, and be able to translate what you've learned outside of security. Visualization concepts are made even more tangible through "makeovers" of security dashboards that many of you may be familiar with.

Finally, we show you how to apply what you've learned at both a personal and organizational level in **Chapter 12**.

Who Should Read This Book

we

We wrote this book because we've both thoroughly enjoy working with data and wholeheartedly believe that we can make significant progress in improving cybersecurity if we take the time to understand how to ask the right questions, perform accurate and reproducible analyses on data, and communicate the results in the most compelling ways possible.

Readers will get the most out of this book if they come to it with some security domain experience and the ability to do basic coding or scripting. If you are already familiar with Python, you can skip the introduction to it in Chapter 2 and can skim through much of Chapter 3. We level the field a bit by introducing and focusing on R, but you would do well to make your way through all the examples and listings that use R throughout the book, as it is an excellent language for modern data science. If you are new to programming, Chapters 2, 3, and 4 will provide enough of an immersive experience to help you see if it's right for you.

(ch 9): "Demystifying Machine Learning"

title of Chap.9 differs from tracking grid

We place emphasis on statistical and machine learning across many chapters and do not recommend skipping any of that content. However, you *can* hold off on Chapter 9 (machine learning) until the very end, as it will not detract significantly from the flow of the book.

If you know databases well, you need only review the use cases in Chapter 8 to ensure you're thinking about all the ways you can use modern and specialized databases in security use cases.

Unlike many books that discuss dashboards, the only requirements for Chapter 10 are Microsoft Excel or OpenOffice Calc, as we made no assumptions about the types of tools and restrictions you have to work within your organization. You can also save Chapter 11 for future reading if you have no desire to build interactive visualizations.

with in

In short, though we are writing to Information Technology and Information Security professionals, students, consultants, and anyone looking for more about the how-to of analyzing data and making it understandable for protecting networks will find what they need in this book.

make sentence case

Tools You Will Need

Everything you need to follow along with the exercises is freely available:

em dash x9 (after lead-ins for book consistency)

- **The R project** (`http://www.r-project.org`): Most of the examples are written in R, and with the wide range of community developed packages like ggplot2 (`http://ggplot2.org`) almost anything is possible.
- **RStudio** (`http://www.rstudio.com/`): It will be *much* easier to get to know R and run the examples if you use the RStudio IDE.
- **Python** (`http://www.python.org`): A few of the examples leverage Python and with add-on packages like pandas (`http://pandas.pydata.org`) makes this a very powerful platform.
- **Sublime Text** (`http://www.sublimetext.com/`): This, or another robust text editor, will come in very handy especially when working with HTML/CSS/JavaScript examples.
- **D3.js** (`http://d3js.org/`): Grabbing a copy of D3 and giving the basis a quick read through ahead of **Chapter 11** will help you work through the examples in that chapter a bit faster.

remove bf

basics / ?

- **Git** (`http://git-scm.com/`): You'll be asked to use git to download data at various points in the book, so installing it now will save you some time later.
- **MongoDB** (`http://www.mongodb.org/`): MongoDB is used in **Chapter 8**, so getting it set up early will make those examples less cumbersome.

not bold

not bold

- **Redis** (`http://redis.io/`): This, too, is used in some examples in **Chapter 8**.
- **Tableau Public** (`http://www.tableausoftware.com/`): If you intend to work with the survey data in *Chapter 11*, having a copy of Tableau Public will be useful.

not italic

Additionally, all of the code, examples, and data used in this book are available through the companion website for this book (`www.wiley.com/go/datadrivensecurity`).

We recommend using Linux or Mac OS, but all of the examples should work fine on modern flavors of Microsoft Windows as well.

What's on the Website

As mentioned earlier, you'll want to check out the companion website `www.wiley.com/go/datadrivensecurity` for the book, which has the full source code for all code listings, the data files used in the examples, and any supporting documents (such as Microsoft Excel files).

The Journey Begins!

You have everything you need to start down the path to *Data-Driven Security*. We hope your journey will be filled with new insights and discoveries and are confident you'll be able to improve your security posture if you successfully apply the principles you're about to learn.

hyphen