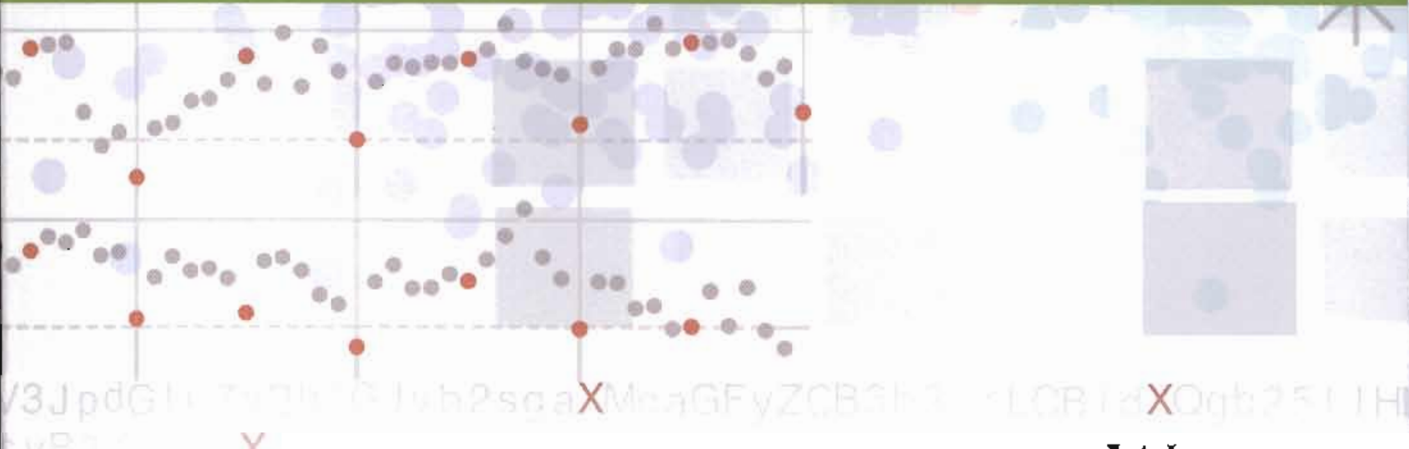
JAY JACOBS + BOB RUDIS

Data-Driven Security

Analysis, Visualization and Dashboards

BIBLIOTHEQUE DU CERIST



WILEY

Data-Driven **Security**

Analysis, Visualization and Dashboards

JAY JACOBS + BOB RUDIS

WILEY

BIBLIOTHEQUE DU CERIST

Contents

Introduction xv

Chapter 1 • The Journey to Data-Driven Security 1

A Brief History of Learning from Data 2

Nineteenth Century Data Analysis 2

Twentieth Century Data Analysis 3

Twenty-First Century Data Analysis 4

Gathering Data Analysis Skills 5

Domain Expertise 6

Programming Skills 8

Data Management 10

Statistics 12

Visualization (a.k.a. Communication) 14

Combining the Skills 15

Centering on a Question 16

Creating a Good Research Question 17

Exploratory Data Analysis 18

Summary 18

Recommended Reading 19

Chapter 2 • Building Your Analytics Toolbox: A Primer on Using R and Python for Security Analysis 21

Why Python? Why R? And Why Both? 22

Why Python? 23

Why R? 23

Why Both? 24

Jumpstarting Your Python Analytics with Canopy 24

Understanding the Python Data Analysis and Visualization Ecosystem 25

Setting Up Your R Environment 29

Introducing Data Frames 33

Organizing Analyses 36

Summary 37

Recommended Reading 38

Chapter 3 • Learning the "Hello World" of Security Data Analysis 39

Solving a Problem 40

Getting Data 41

Reading In Data 43

Exploring Data 47

BIBLIOTHEQUE DU CERIST

Homing In on a Question 58
Summary 70
Recommended Reading 70

Chapter 4 • Performing Exploratory Security Data Analysis 71

Dissecting the IP Address 73
Representing IP Addresses 73
Segmenting and Grouping IP Addresses 75
Locating IP Addresses 77
Augmenting IP Address Data 80
Association/Correlation, Causation, and Security Operations Center Analysts Gone Rogue 86
Mapping Outside the Continents 90
Visualizing the ZeuS Botnet 92
Visualizing Your Firewall Data 98
Summary 100
Recommended Reading 101

Chapter 5 • From Maps to Regression 103

Simplifying Maps 105
How Many ZeroAccess Infections per Country? 108
Changing the Scope of Your Data 111
The Potwin Effect 113
Is This Weird? 117
Counting in Counties 120
Moving Down to Counties 122
Introducing Linear Regression 125
Understanding Common Pitfalls in Regression Analysis 130
Regression on ZeroAccess Infections 131
Summary 136
Recommended Reading 136

Chapter 6 • Visualizing Security Data 137

Why Visualize? 138
Unraveling Visual Perception 139
Understanding the Components of Visual Communications 144
Avoiding the Third Dimension 144
Using Color 146
Putting It All Together 148
Communicating Distributions 154
Visualizing Time Series 156
Experiment on Your Own 157
Turning Your Data into a Movie Star 158
Summary 159
Recommended Reading 160

BIBLIOTHEQUE DU CERIST

Chapter 7 • Learning from Security Breaches 161
Setting Up the Research 162
Considerations in a Data Collection Framework 164
Aiming for Objective Answers *164*
Limiting Possible Answers *164*
Allowing "Other," and "Unknown" Options *164*
Avoiding Conflation and Merging the Minutiae *165*
An Introduction to VERIS 166
Incident Tracking *168*
Threat Actor *168*
Threat Actions *169*
Information Assets *173*
Attributes *173*
Discovery/Response *176*
Impact *176*
Victim *177*
Indicators *179*
Extending VERIS with Plus *179*
Seeing VERIS in Action 179
Working with VCDB Data 181
Getting the Most Out of VERIS Data *185*
Summary 189
Recommended Reading 189

Chapter 8 • Breaking Up with Your Relational Database 191
Realizing the Container Has Constraints 195
Constrained by Schema *196*
Constrained by Storage *198*
Constrained by RAM *199*
Constrained by Data *200*
Exploring Alternative Data Stores 200
BerkeleyDB *201*
Redis *203*
Hive *207*
MongoDB *210*
Special Purpose Databases *214*
Summary 215
Recommended Reading 216

Chapter 9 • Demystifying Machine Learning 217
Detecting Malware 218
Developing a Machine Learning Algorithm *220*
Validating the Algorithm *221*
Implementing the Algorithm *222*

BIBLIOTHEQUE DU CERIST

Benefiting from Machine Learning . . . 226
Answering Questions with Machine Learning . . . *226*
Measuring Good Performance. . . . *227*
Selecting Features . . . *228*
Validating Your Model . . . *230*
Specific Learning Methods. . . . 230
Supervised . . . *231*
Unsupervised. . . . *234*
Hands On: Clustering Breach Data. . . . 236
Multidimensional Scaling on Victim Industries . . . *238*
Hierarchical Clustering on Victim Industries . . . *240*
Summary . . . 242
Recommended Reading . . . 243

Chapter 10 • Designing Effective Security Dashboards. . . . 245
What Is a Dashboard, Anyway? . . . 246
A Dashboard Is Not an Automobile . . . *246*
A Dashboard Is Not a Report . . . *248*
A Dashboard Is Not a Moving Van . . . *251*
A Dashboard Is Not an Art Show. . . . *253*
Communicating and Managing "Security" through Dashboards . . . 258
Lending a Hand to Handlers. . . . *258*
Raising Dashboard Awareness . . . *260*
The Devil (and Incident Response Delays) Is in the Details . . . *262*
Projecting "Security" . . . *263*
Summary . . . 267
Recommended Reading . . . 267

Chapter 11 • Building Interactive Security Visualizations . . . 269
Moving from Static to Interactive . . . 270
Interaction for Augmentation . . . *271*
Interaction for Exploration . . . *274*
Interaction for Illumination . . . *276*
Developing Interactive Visualizations . . . 281
Building Interactive Dashboards with Tableau . . . *281*
Building Browser-Based Visualizations with D3. . . . *284*
Summary . . . 294
Recommended Reading . . . 295

Chapter 12 • Moving Toward Data-Driven Security . . . 297
Moving Yourself toward Data-Driven Security . . . 298
The Hacker . . . *299*
The Statistician . . . *302*
The Security Domain Expert . . . *302*
The Danger Zone . . . *303*

BIBLIOTHEQUE DU CERIST

Moving Your Organization toward Data-Driven Security . . . 303
Ask Questions That Have Objective Answers . . . *304*
Find and Collect Relevant Data . . . *304*
Learn through Iteration . . . *305*
Find Statistics . . . *306*
Summary . . . 308
Recommended Reading . . . 308

Appendix A • Resources and Tools . . . 309

Appendix B • References . . . 313

Index • . . . 321

BIBLIOTHEQUE DU CERIST