

BRIEF CONTENTS

| | |
|---|------|
| Foreword | xvii |
| Acknowledgments | xix |
| Introduction | xxi |
| Chapter 1: Basic Static Malware Analysis | 1 |
| Chapter 2: Beyond Basic Static Analysis: x86 Disassembly | 11 |
| Chapter 3: A Brief Introduction to Dynamic Analysis | 25 |
| Chapter 4: Identifying Attack Campaigns Using Malware Networks | 35 |
| Chapter 5: Shared Code Analysis | 59 |
| Chapter 6: Understanding Machine Learning–Based Malware Detectors | 89 |
| Chapter 7: Evaluating Malware Detection Systems | 119 |
| Chapter 8: Building Machine Learning Detectors | 127 |
| Chapter 9: Visualizing Malware Trends | 155 |
| Chapter 10: Deep Learning Basics | 175 |
| Chapter 11: Building a Neural Network Malware Detector with Keras | 199 |
| Chapter 12: Becoming a Data Scientist | 215 |
| Appendix: An Overview of Datasets and Tools | 221 |
| Index | 233 |

CONTENTS IN DETAIL

| | |
|-------------------------------|-------------|
| FOREWORD by Anup Ghosh | xvii |
|-------------------------------|-------------|

| | |
|------------------------|------------|
| ACKNOWLEDGMENTS | xix |
|------------------------|------------|

| | |
|---------------------|------------|
| INTRODUCTION | xxi |
|---------------------|------------|

| | |
|---|-------|
| What Is Data Science? | xxii |
| Why Data Science Matters for Security | xxii |
| Applying Data Science to Malware | xxiii |
| Who Should Read This Book? | xxiv |
| About This Book. | xxiv |
| How to Use the Sample Code and Data | xxv |

| | |
|--------------------------------------|----------|
| 1 | |
| BASIC STATIC MALWARE ANALYSIS | 1 |

| | |
|---|----|
| The Microsoft Windows Portable Executable Format. | 2 |
| The PE Header | 3 |
| The Optional Header | 3 |
| Section Headers | 4 |
| Dissecting the PE Format Using pefile | 5 |
| Examining Malware Images | 7 |
| Examining Malware Strings. | 8 |
| Using the strings Program | 8 |
| Analyzing Your strings Dump | 9 |
| Summary | 10 |

| | |
|--|-----------|
| 2 | |
| BEYOND BASIC STATIC ANALYSIS: X86 DISASSEMBLY | 11 |

| | |
|---|----|
| Disassembly Methods | 12 |
| Basics of x86 Assembly Language | 12 |
| CPU Registers | 13 |
| Arithmetic Instructions | 15 |
| Data Movement Instructions | 15 |
| Disassembling ircbot.exe Using pefile and capstone. | 20 |
| Factors That Limit Static Analysis | 21 |
| Packing | 21 |
| Resource Obfuscation | 22 |
| Anti-disassembly Techniques. | 22 |
| Dynamically Downloaded Data | 22 |
| Summary | 23 |

| | | |
|----------|--|-----------|
| 3 | A BRIEF INTRODUCTION TO DYNAMIC ANALYSIS | 25 |
| | Why Use Dynamic Analysis? | 26 |
| | Dynamic Analysis for Malware Data Science. | 26 |
| | Basic Tools for Dynamic Analysis | 27 |
| | Typical Malware Behaviors | 27 |
| | Loading a File on malwr.com | 27 |
| | Analyzing Results on malwr.com. | 28 |
| | Limitations of Basic Dynamic Analysis | 33 |
| | Summary | 34 |
| | | |
| 4 | IDENTIFYING ATTACK CAMPAIGNS USING MALWARE NETWORKS | 35 |
| | Nodes and Edges | 37 |
| | Bipartite Networks | 37 |
| | Visualizing Malware Networks | 39 |
| | The Distortion Problem | 39 |
| | Force-Directed Algorithms | 40 |
| | Building Networks with NetworkX | 40 |
| | Adding Nodes and Edges. | 41 |
| | Adding Attributes | 42 |
| | Saving Networks to Disk | 42 |
| | Network Visualization with GraphViz. | 43 |
| | Using Parameters to Adjust Networks | 44 |
| | The GraphViz Command Line Tools | 44 |
| | Adding Visual Attributes to Nodes and Edges | 48 |
| | Building Malware Networks | 51 |
| | Building a Shared Image Relationship Network | 54 |
| | Summary | 58 |
| | | |
| 5 | SHARED CODE ANALYSIS | 59 |
| | Preparing Samples for Comparison by Extracting Features | 62 |
| | How Bag of Features Models Work. | 62 |
| | What are N-Grams? | 63 |
| | Using the Jaccard Index to Quantify Similarity | 64 |
| | Using Similarity Matrices to Evaluate Malware Shared Code Estimation Methods | 66 |
| | Instruction Sequence–Based Similarity | 67 |
| | Strings-Based Similarity | 70 |
| | Import Address Table–Based Similarity | 71 |
| | Dynamic API Call–Based Similarity | 72 |
| | Building a Similarity Graph | 73 |
| | Scaling Similarity Comparisons | 77 |
| | Minhash in a Nutshell | 77 |
| | Minhash in Depth | 78 |
| | Building a Persistent Malware Similarity Search System | 79 |
| | Running the Similarity Search System | 85 |
| | Summary | 87 |

6 UNDERSTANDING MACHINE LEARNING–BASED MALWARE DETECTORS 89

- Steps for Building a Machine Learning–Based Detector 90
 - Gathering Training Examples 91
 - Extracting Features 91
 - Designing Good Features. 92
 - Training Machine Learning Systems. 92
 - Testing Machine Learning Systems 93
- Understanding Feature Spaces and Decision Boundaries 93
- What Makes Models Good or Bad: Overfitting and Underfitting 98
- Major Types of Machine Learning Algorithms 101
 - Logistic Regression 102
 - K-Nearest Neighbors. 105
 - Decision Trees 109
 - Random Forest 115
- Summary 117

7 EVALUATING MALWARE DETECTION SYSTEMS 119

- Four Possible Detection Outcomes 120
 - True and False Positive Rates 120
 - Relationship Between True and False Positive Rates 121
 - ROC Curves. 123
- Considering Base Rates in Your Evaluation 124
 - How Base Rate Affects Precision 124
 - Estimating Precision in a Deployment Environment 125
- Summary 126

8 BUILDING MACHINE LEARNING DETECTORS 127

- Terminology and Concepts 128
- Building a Toy Decision Tree–Based Detector. 129
 - Training Your Decision Tree Classifier 130
 - Visualizing the Decision Tree 131
 - Complete Sample Code. 133
- Building Real-World Machine Learning Detectors with sklearn 134
 - Real-World Feature Extraction 134
 - Why You Can’t Use All Possible Features. 137
 - Using the Hashing Trick to Compress Features 138
- Building an Industrial-Strength Detector 141
 - Extracting Features 141
 - Training the Detector. 142
 - Running the Detector on New Binaries. 144
 - What We’ve Implemented So Far 144
- Evaluating Your Detector’s Performance 146
 - Using ROC Curves to Evaluate Detector Efficacy 147
 - Computing ROC Curves. 147
 - Splitting Data into Training and Test Sets 148

| | |
|-----------------------------------|-----|
| Computing the ROC Curve | 149 |
| Cross-Validation | 150 |
| Next Steps | 153 |
| Summary | 154 |

9 VISUALIZING MALWARE TRENDS 155

| | |
|--|-----|
| Why Visualizing Malware Data Is Important | 156 |
| Understanding Our Malware Dataset | 158 |
| Loading Data into pandas | 158 |
| Working with a pandas DataFrame | 159 |
| Filtering Data Using Conditions | 161 |
| Using matplotlib to Visualize Data | 162 |
| Plotting the Relationship Between Malware Size and Detection | 162 |
| Plotting Ransomware Detection Rates | 164 |
| Plotting Ransomware and Worm Detection Rates | 165 |
| Using seaborn to Visualize Data | 168 |
| Plotting the Distribution of Antivirus Detections | 169 |
| Creating a Violin Plot | 172 |
| Summary | 174 |

10 DEEP LEARNING BASICS 175

| | |
|--|-----|
| What Is Deep Learning? | 176 |
| How Neural Networks Work | 177 |
| Anatomy of a Neuron | 177 |
| A Network of Neurons | 180 |
| Universal Approximation Theorem | 181 |
| Building Your Own Neural Network | 182 |
| Adding Another Neuron to the Network | 186 |
| Automatic Feature Generation | 188 |
| Training Neural Networks | 189 |
| Using Backpropagation to Optimize a Neural Network | 190 |
| Path Explosion | 192 |
| Vanishing Gradient | 192 |
| Types of Neural Networks | 193 |
| Feed-Forward Neural Network | 193 |
| Convolutional Neural Network | 193 |
| Autoencoder Neural Network | 194 |
| Generative Adversarial Network | 195 |
| Recurrent Neural Network | 196 |
| ResNet | 196 |
| Summary | 197 |

11 BUILDING A NEURAL NETWORK MALWARE DETECTOR WITH KERAS 199

| | |
|---|-----|
| Defining a Model's Architecture | 200 |
| Compiling the Model | 202 |

| | |
|---|-----|
| Training the Model | 203 |
| Extracting Features | 203 |
| Creating a Data Generator | 204 |
| Incorporating Validation Data | 207 |
| Saving and Loading the Model | 209 |
| Evaluating the Model | 209 |
| Enhancing the Model Training Process with Callbacks | 211 |
| Using a Built-in Callback | 212 |
| Using a Custom Callback | 213 |
| Summary | 214 |

12

| | |
|--|------------|
| BECOMING A DATA SCIENTIST | 215 |
| Paths to Becoming a Security Data Scientist | 216 |
| A Day in the Life of a Security Data Scientist | 216 |
| Traits of an Effective Security Data Scientist | 218 |
| Open-Mindedness | 218 |
| Boundless Curiosity | 218 |
| Obsession with Results | 219 |
| Skepticism of Results | 219 |
| Where to Go from Here | 219 |

| | |
|--|------------|
| APPENDIX | |
| AN OVERVIEW OF DATASETS AND TOOLS | 221 |
| Overview of Datasets | 222 |
| Chapter 1: Basic Static Malware Analysis | 222 |
| Chapter 2: Beyond Basic Static Analysis: x86 Disassembly | 222 |
| Chapter 3: A Brief Introduction to Dynamic Analysis | 222 |
| Chapter 4: Identifying Attack Campaigns Using Malware Networks | 222 |
| Chapter 5: Shared Code Analysis | 223 |
| Chapter 6: Understanding Machine Learning–Based Malware Detectors and Chapter 7: Evaluating Malware Detection Systems | 223 |
| Chapter 8: Building Machine Learning Detectors | 224 |
| Chapter 9: Visualizing Malware Trends | 224 |
| Chapter 10: Deep Learning Basics | 224 |
| Chapter 11: Building a Neural Network Malware Detector with Keras | 224 |
| Chapter 12: Becoming a Data Scientist | 224 |
| Tool Implementation Guide | 225 |
| Shared Hostname Network Visualization | 225 |
| Shared Image Network Visualization | 226 |
| Malware Similarity Visualization | 227 |
| Malware Similarity Search System | 229 |
| Machine Learning Malware Detection System | 230 |

| | |
|--------------|------------|
| INDEX | 233 |
|--------------|------------|