

CONTENTS IN DETAIL

ACKNOWLEDGMENTS	xxi
------------------------	------------

INTRODUCTION	xxiii
---------------------	--------------

Why I Wrote This Book	xxiv
What You'll Learn	xxv
What You'll Need	xxviii

PART I: SOURCES AND DATASETS **1**

1	
PROTECTING SOURCES AND YOURSELF	3

Safely Communicating with Sources	4
Working with Public Data	5
Protecting Sensitive Information	5
Minimizing the Digital Trail	5
Working with Hackers and Whistleblowers	6
Secure Storage for Datasets	7
Low-Sensitivity Datasets	7
Medium-Sensitivity Datasets	8
High-Sensitivity Datasets	9
Authenticating Datasets	10
The AFLDS Dataset	11
The WikiLeaks Twitter Group Chat	12
Redaction	14
What Data to Publish	15
What to Redact	15
Making Requests for Comment	17
Password Managers	18
Disk Encryption	21
Exercise 1-1: Encrypt Your Internal Disk	22
Windows	22
macOS	24
Linux	24
Exercise 1-2: Encrypt a USB Disk	25
Windows	26
macOS	28
Linux	28
Protecting Yourself from Malicious Documents	28
Exercise 1-3: Install and Use Dangerzone	30
Summary	32

2		
ACQUIRING DATASETS		33
The End of WikiLeaks		34
Distributed Denial of Secrets		35
Downloading Datasets with BitTorrent		37
The Origins of BlueLeaks		38
Exercise 2-1: Download the BlueLeaks Dataset.		39
Communicating with Encrypted Messaging Apps		39
Exercise 2-2: Install and Practice Using Signal		41
Encrypting Messages with PGP		42
Staying Anonymous Online with Tor and OnionShare		42
Exercise 2-3: Play with Tor and OnionShare		46
Communicating with My Tea Party Patriots Source		47
Other Options for Acquiring Datasets from Sources		47
Encrypted USB Drives		48
Virtual Private Servers		49
Whistleblower Submission Systems		49
Summary		51

PART II: TOOLS OF THE TRADE **53**

3		
THE COMMAND LINE INTERFACE		55
Introducing the Command Line		56
The Shell		56
Users and Paths		57
User Privileges		58
Exercise 3-1: Install Ubuntu in Windows		59
Basic Command Line Usage		62
Opening a Terminal		62
Clearing Your Screen and Exiting the Shell		63
Exploring Files and Directories		63
Navigating Relative and Absolute Paths		65
Changing Directories		65
Using the help Argument		66
Accessing Man Pages		67
Tips for Navigating the Terminal		67
Entering Commands with Tab Completion		67
Editing Commands		68
Dealing with Spaces in Filenames		69
Using Single Quotes Around Double Quotes		70
Installing and Uninstalling Software with Package Managers		70
Exercise 3-2: Manage Packages with Homebrew on macOS.		72
Exercise 3-3: Manage Packages with apt on Windows or Linux		74
Exercise 3-4: Practice Using the Command Line with cURL		76
Download a Web Page with cURL		76
Save a Web Page to a File		77
Text Files vs. Binary Files		77

Exercise 3-5: Install the VS Code Text Editor	78
Exercise 3-6: Write Your First Shell Script	80
Navigate to Your USB Disk	80
Create an Exercises Folder	81
Open a VS Code Workspace	81
Write the Shell Script	82
Run the Shell Script	83
Exercise 3-7: Clone the Book's GitHub Repository	84
Summary	85

4 EXPLORING DATASETS IN THE TERMINAL 87

Introducing for Loops	88
Exercise 4-1: Unzip the BlueLeaks Dataset	89
Unzip Files on macOS or Linux	89
Unzip Files on Windows	92
Organize Your Files	93
How the Hacker Obtained the BlueLeaks Data	94
Exercise 4-2: Explore BlueLeaks on the Command Line	96
Calculate How Much Disk Space Folders Use	96
Use Pipes and Sort Output	97
Create an Inventory of Filenames in a Dataset	99
Count the Files in a Dataset	100
Exercise 4-3: Find Revelations in BlueLeaks with grep	100
Filter for Documents Mentioning Antifa	100
Filter for Certain Types of Files	102
Use grep with Regular Expressions	102
Search Files in Bulk with grep	103
Encrypted Data in the BlueLeaks Dataset	104
Data Analysis with Servers in the Cloud	106
Exercise 4-4: Set Up a VPS	108
Generate an SSH Key	108
Add Your Public Key to the Cloud Provider	109
Create a VPS	110
SSH into Your Server	111
Start a Byobu Session	111
Install Updates	112
Exercise 4-5: Explore the Oath Keepers Dataset Remotely	113
Summary	117

5 DOCKER, ALEPH, AND MAKING DATASETS SEARCHABLE 119

Introducing Docker and Linux Containers	120
Exercise 5-1: Initialize Docker Desktop on Windows and macOS	121
Exercise 5-2: Initialize Docker Engine on Linux	122
Running Containers with Docker	123
Running an Ubuntu Container	123
Listing and Killing Containers	124
Mounting and Removing Volumes	125
Passing Environment Variables	129

Running Server Software	129
Freeing Up Disk Space	132
Exercise 5-3: Run a WordPress Site with Docker Compose	132
Make a docker-compose.yaml File	132
Start Your WordPress Site	134
Introducing Aleph	135
Exercise 5-4: Run Aleph Locally in Linux Containers	136
Using Aleph’s Web and Command Line Interfaces	138
Indexing Data in Aleph	139
Exercise 5-5: Index a BlueLeaks Folder in Aleph	140
Mount Your Datasets into the Aleph Shell	140
Index the icefishx Folder	141
Check Indexing Status	142
Explore BlueLeaks with Aleph	144
Additional Aleph Features	145
Dedicated Aleph Servers	147
Summary	148

6 READING OTHER PEOPLE’S EMAIL 149

The Email Protocol and Message Structure	150
File Formats for Email Dumps	151
EML Files	152
MBOX Files	152
PST Outlook Data Files	152
Exercise 6-1: Download Email Dumps from Three Datasets	153
The Nauru Police Force Dataset	153
The Oath Keepers Dataset	153
The Heritage Foundation Dataset	154
Researching Email Dumps with Thunderbird	154
Exercise 6-2: Configure Thunderbird for Email Dumps	155
Reading Individual EML Files with Thunderbird	156
Exercise 6-3: Import the Nauru Police Force EML Email Dump	157
Searching Email in Thunderbird	159
Quick Filter Searches	159
The Search Messages Dialog	159
Exercise 6-4: Import the Oath Keepers MBOX Email Dump	160
Exercise 6-5: Import the Heritage Foundation PST Email Dump	161
Other Tools for Researching Email Dumps	163
Microsoft Outlook	163
Aleph	165
Summary	166

7

AN INTRODUCTION TO PYTHON**169**

Exercise 7-1: Install Python	170
Windows	170
Linux	170
macOS	170
Exercise 7-2: Write Your First Python Script	171
Python Basics	172
The Interactive Python Interpreter	172
Comments	173
Math with Python	173
Strings	175
Exercise 7-3: Write a Python Script with Variables, Math, and Strings	176
Lists and Loops	178
Defining and Printing Lists	178
Running for Loops	180
Control Flow	182
Comparison Operators	182
if Statements	183
Nested Code Blocks	185
Searching Lists	185
Logical Operators	186
Exception Handling	187
Exercise 7-4: Practice Loops and Control Flow	190
Functions	192
The def Keyword	192
Default Arguments	193
Return Values	194
Docstrings	196
Exercise 7-5: Practice Writing Functions	196
Summary	198

8

WORKING WITH DATA IN PYTHON**199**

Modules	200
Python Script Template	201
Exercise 8-1: Traverse the Files in BlueLeaks	202
List the Filenames in a Folder	202
Count the Files and Folders in a Folder	203
Traverse Folders with os.walk()	205
Exercise 8-2: Find the Largest Files in BlueLeaks	207
Third-Party Modules	208
Exercise 8-3: Practice Command Line Arguments with Click	210
Avoiding Hardcoding with Command Line Arguments	212
Exercise 8-4: Find the Largest Files in Any Dataset	213

Dictionaries	214
Defining Dictionaries	214
Getting and Setting Values	215
Navigating Dictionaries and Lists in the Conti Chat Logs	216
Exploring Dictionaries and Lists Full of Data in Python	216
Selecting Values in Dictionaries and Lists	219
Analyzing Data Stored in Dictionaries and Lists	220
Exercise 8-5: Map Out the CSVs in BlueLeaks.	223
Accept a Command Line Argument	224
Loop Through the BlueLeaks Folders.	225
Fill Up the Dictionary	226
Display the Output	227
Reading and Writing Files.	229
Opening Files.	229
Writing Lines to a File	230
Reading Lines from a File	230
Exercise 8-6: Practice Reading and Writing Files	231
Summary	233

PART IV: STRUCTURED DATA

235

9

BLUELEAKS, BLACK LIVES MATTER, AND THE CSV FILE FORMAT

237

Installing Spreadsheet Software	238
Introducing the CSV File Format	238
Exploring CSV Files with Spreadsheet Software and Text Editors	240
My BlueLeaks Investigation	243
Focusing on a Fusion Center.	243
Introducing NCRIC	244
Investigating a SAR	244
Reading and Writing CSV Files in Python	248
Exercise 9-1: Make BlueLeaks CSVs More Readable.	250
Accept the CSV Path as an Argument	250
Loop Through the CSV Rows.	251
Display CSV Fields on Separate Lines	252
How to Read Bulk Email from Fusion Centers.	254
Lists of Black Lives Matter Demonstrations.	254
“Intelligence” Memos from the FBI and DHS	259
A Brief HTML Primer	260
Exercise 9-2: Make Bulk Email Readable	262
Accept the Command Line Arguments	262
Create the Output Folder	263
Define the Filename for Each Row	264
Write the HTML Version of Each Bulk Email	265
Discovering the Names and URLs of BlueLeaks Sites	270
Exercise 9-3: Make a CSV of BlueLeaks Sites	271
Open a CSV for Writing	272

Find All the Company.csv Files	273
Add BlueLeaks Sites to the CSV	274
Summary	276

10 BLUELEAKS EXPLORER 277

Undiscovered Revelations in BlueLeaks	278
Exercise 10-1: Install BlueLeaks Explorer	279
Create the Docker Compose Configuration File.	279
Bring Up the Containers.	280
Initialize the Databases	280
The Structure of NCRIC.	282
Exploring Tables and Relationships	282
Searching for Keywords.	285
Building Your Own BlueLeaks Structure	286
Defining the JRIC Structure	286
Showing Useful Fields	288
Changing Field Types	290
Adding JRIC’s Leads Table	292
Building a Relationship	293
Verifying BlueLeaks Data	295
Exercise 10-2: Finish Building the Structure for JRIC	296
The Technology Behind BlueLeaks Explorer	297
The Backend.	298
The Frontend	298
Summary	299

11 PARLER, THE JANUARY 6 INSURRECTION, AND THE JSON FILE FORMAT 301

The Origins of the Parler Dataset	302
How the Parler Videos Were Archived.	302
The Dataset’s Impact on Trump’s Second Impeachment.	303
Exercise 11-1: Download and Extract Parler Video Metadata	304
Download the Metadata	304
Uncompress and Download Individual Parler Videos	306
Extract Parler Metadata	308
The JSON File Format.	309
Understanding JSON Syntax	309
Parsing JSON with Python	312
Handling Exceptions with JSON	314
Tools for Exploring JSON Data	316
Counting Videos with GPS Coordinates Using grep.	316
Formatting and Searching Data with the jq Command.	317
Exercise 11-2: Write a Script to Filter for Videos with GPS from January 6, 2021	318
Accept the Parler Metadata Path as an Argument	319
Loop Through Parler Metadata Files.	320
Filter for Videos with GPS Coordinates	321
Filter for Videos from January 6, 2021	323

Working with GPS Coordinates	324
Searching by Latitude and Longitude	324
Converting Between GPS Coordinate Formats	326
Calculating GPS Distance in Python.	327
Finding the Center of Washington, DC	329
Exercise 11-3: Update the Script to Filter for Insurrection Videos.	330
Plotting GPS Coordinates on a Map with simplekml	333
Exercise 11-4: Create KML Files to Visualize Location Data	335
Create a KML File for All Videos with GPS Coordinates	336
Create KML Files for Videos from January 6, 2021	339
Visualizing Location Data with Google Earth	341
Viewing Metadata with ExifTool	344
Summary	346

12

EPIK FAIL, EXTREMISM RESEARCH, AND SQL DATABASES 347

The Structure of SQL Databases	348
Relational Databases	349
Clients and Servers	350
Tables, Columns, and Types	351
Exercise 12-1: Create and Test a MySQL Server	
Using Docker and Adminer	352
Run the Server	352
Connect to the Database with Adminer	353
Create a Test Database	354
Exercise 12-2: Query Your SQL Database	355
INSERT Statements	356
SELECT Statements	357
JOIN Clauses	362
UPDATE Statements	365
DELETE Statements.	366
Introducing the MySQL Command Line Client	366
Exercise 12-3: Install and Test the Command Line MySQL Client	367
MySQL-Specific Queries	368
The History of Epik.	370
The Epik Hack	371
Epik’s WHOIS Data.	373
Exercise 12-4: Download and Extract Part of the Epik Dataset	375
Exercise 12-5: Import Epik Data into MySQL	375
Create a Database for api_system.	376
Import api_system Data	376
Exploring Epik’s SQL Database	378
The domain Table	378
The privacy Table	380
The hosting and hosting_server Tables	382
Working with Epik Data in the Cloud	384
Summary	386

PART V: CASE STUDIES

387

13

PANDEMIC PROFITEERS AND COVID-19 DISINFORMATION 389

The Origins of AFLDS	391
The Cadence Health and Ravkoo Datasets	393
Extracting the Data into an Encrypted File Container	394
Analyzing the Data with Command Line Tools.	395
Creating a Single Spreadsheet of Patients	402
Calculating Revenue from Prescriptions Filled by Ravkoo.	405
Finding the Price and Quantity of Drugs Sold	406
Categorizing Prescription Data by Drug.	408
A Deeper Look at the Cadence Health Patient Data	411
Finding Cadence’s Partners	411
Searching for Patients by City.	414
Searching for Patients by Age.	418
Authenticating the Data.	421
The Aftermath	423
HIPAA’s Breach Notification Rule.	424
Congressional Investigation	424
Simone Gold’s New Business Venture	425
Scandal and Infighting at AFLDS.	425
Summary	426

14

NEO-NAZIS AND THEIR CHATROOMS 427

How Antifascists Infiltrated Neo-Nazi Discord Servers.	428
Analyzing Leaked Chat Logs	429
Making JSON Files Readable	430
Exploring Objects, Keys, and Values with jq.	431
Converting Timestamps	436
Finding Usernames	436
The Discord History Tracker.	438
A Script to Search the JSON Files	440
My Discord Analysis Code	444
Designing the SQL Database	445
Importing Chat Logs into the SQL Database	448
Building the Web Interface.	454
Using Discord Analysis to Find Revelations.	459
The Pony Power Discord Server	462
The Launch of DiscordLeaks.	466
The Aftermath	466
The Lawsuit Against Unite the Right	467
The Patriot Front Chat Logs.	468
Summary	469

AFTERWORD **471**

A
SOLUTIONS TO COMMON WSL PROBLEMS **473**

Understanding WSL's Linux Filesystem 474
The Disk Performance Problem 476
Solving the Disk Performance Problem 477
 Storing Only Active Datasets in Linux 477
 Storing Your Linux Filesystem on a USB Disk 477
Next Steps 482

B
SCRAPING THE WEB **483**

Legal Considerations 484
HTTP Requests 485
Scraping Techniques 486
 Loading Pages with HTTPX 486
 Parsing HTML with BeautifulSoup 491
 Automating Web Browsers with Selenium 496
Next Steps 501

INDEX **503**