
Table of Contents

Preface.....	xi
1. Introduction.....	1
Overview	2
Data Science Is OSEMN	2
Obtaining Data	2
Scrubbing Data	3
Exploring Data	3
Modeling Data	3
Interpreting Data	4
Intermezzo Chapters	4
What Is the Command Line?	5
Why Data Science at the Command Line?	7
The Command Line Is Agile	7
The Command Line Is Augmenting	7
The Command Line Is Scalable	8
The Command Line Is Extensible	8
The Command Line Is Ubiquitous	9
A Real-World Use Case	9
Further Reading	12
2. Getting Started.....	13
Overview	13
Setting Up Your Data Science Toolbox	13
Step 1: Download and Install VirtualBox	14
Step 2: Download and Install Vagrant	14
Step 3: Download and Start the Data Science Toolbox	15
Step 4: Log In (on Linux and Mac OS X)	16

Step 4: Log In (on Microsoft Windows)	17
Step 5: Shut Down or Start Anew	17
Essential Concepts and Tools	17
The Environment	18
Executing a Command-Line Tool	19
Five Types of Command-Line Tools	20
Combining Command-Line Tools	23
Redirecting Input and Output	24
Working with Files	24
Help!	25
Further Reading	27
3. Obtaining Data.....	29
Overview	29
Copying Local Files to the Data Science Toolbox	30
Local Version of Data Science Toolbox	30
Remote Version of Data Science Toolbox	30
Decompressing Files	31
Converting Microsoft Excel Spreadsheets	32
Querying Relational Databases	34
Downloading from the Internet	35
Calling Web APIs	37
Further Reading	39
4. Creating Reusable Command-Line Tools.....	41
Overview	42
Converting One-Liners into Shell Scripts	42
Step 1: Copy and Paste	44
Step 2: Add Permission to Execute	45
Step 3: Define Shebang	46
Step 4: Remove Fixed Input	47
Step 5: Parameterize	47
Step 6: Extend Your PATH	48
Creating Command-Line Tools with Python and R	49
Porting the Shell Script	50
Processing Streaming Data from Standard Input	52
Further Reading	53
5. Scrubbing Data.....	55
Overview	56
Common Scrub Operations for Plain Text	56
Filtering Lines	57

Extracting Values	60
Replacing and Deleting Values	62
Working with CSV	62
Bodies and Headers and Columns, Oh My!	62
Performing SQL Queries on CSV	67
Working with HTML/XML and JSON	67
Common Scrub Operations for CSV	72
Extracting and Reordering Columns	72
Filtering Lines	73
Merging Columns	75
Combining Multiple CSV Files	77
Further Reading	80
6. Managing Your Data Workflow.....	81
Overview	82
Introducing Drake	82
Installing Drake	82
Obtain Top Ebooks from Project Gutenberg	84
Every Workflow Starts with a Single Step	85
Well, That Depends	87
Rebuilding Specific Targets	89
Discussion	90
Further Reading	90
7. Exploring Data.....	91
Overview	92
Inspecting Data and Its Properties	92
Header or Not, Here I Come	92
Inspect All the Data	92
Feature Names and Data Types	93
Unique Identifiers, Continuous Variables, and Factors	95
Computing Descriptive Statistics	96
Using csvstat	96
Using R from the Command Line with Rio	99
Creating Visualizations	102
Introducing Gnuplot and feedgnuplot	102
Introducing ggplot2	104
Histograms	107
Bar Plots	108
Density Plots	110
Box Plots	111
Scatter Plots	112

Line Graphs	113
Summary	114
Further Reading	114
8. Parallel Pipelines.....	115
Overview	116
Serial Processing	116
Looping Over Numbers	116
Looping Over Lines	117
Looping Over Files	118
Parallel Processing	119
Introducing GNU Parallel	121
Specifying Input	122
Controlling the Number of Concurrent Jobs	123
Logging and Output	123
Creating Parallel Tools	124
Distributed Processing	125
Get a List of Running AWS EC2 Instances	126
Running Commands on Remote Machines	127
Distributing Local Data Among Remote Machines	128
Processing Files on Remote Machines	129
Discussion	132
Further Reading	133
9. Modeling Data.....	135
Overview	136
More Wine, Please!	136
Dimensionality Reduction with Tapkee	139
Introducing Tapkee	140
Installing Tapkee	140
Linear and Nonlinear Mappings	141
Clustering with Weka	142
Introducing Weka	143
Taming Weka on the Command Line	143
Converting Between CSV and ARFF	147
Comparing Three Clustering Algorithms	147
Regression with SciKit-Learn Laboratory	150
Preparing the Data	150
Running the Experiment	151
Parsing the Results	151
Classification with BigML	153
Creating Balanced Train and Test Data Sets	153

Calling the API	155
Inspecting the Results	155
Conclusion	156
Further Reading	156
10. Conclusion.....	159
Let's Recap	159
Three Pieces of Advice	160
Be Patient	160
Be Creative	161
Be Practical	161
Where to Go from Here?	161
APIs	161
Shell Programming	162
Python, R, and SQL	162
Interpreting Data	162
Getting in Touch	162
A. List of Command-Line Tools.....	165
B. Bibliography.....	183
Index.....	187