Table of Contents

About the Authors ix

Preface xiii

1. Setting the Pace: What Is Bad Data? 1

2. Is It Just Me, or Does This Data Smell Funny? 5

Understand the Data Structure 6

Field Validation 9

Value Validation 10

Physical Interpretation of Simple Statistics 11

Visualization 12

Keyword PPC Example 14

Search Referral Example 19

Recommendation Analysis 21

Time Series Data 24

Conclusion 29

3. Data Intended for Human Consumption, Not Machine Consumption 31

The Data 31

The Problem: Data Formatted for Human Consumption 32

The Arrangement of Data 32

Data Spread Across Multiple Files 37

The Solution: Writing Code 38

Reading Data from an Awkward Format 39

Reading Data Spread Across Several Files 40

Postscript 48

Other Formats 48

Summary 51

4. Bad Data Lurking in Plain Text 53

McCallum, Q. Ethan
Bad data handbook
2013

digitalisiert durch:
IDS Basel Bern

Which Plain Text Encoding? 54
Guessing Text Encoding 58
Normalizing Text 61
Problem: Application-Specific Characters Leaking into Plain Text 63
Text Processing with Python 67
Exercises 68

5. **(Re)Organizing the Web's Data.. 69**
Can You Get That? 70
General Workflow Example 71
robots.txt 72
Identifying the Data Organization Pattern 73
Store Offline Version for Parsing 75
Scrape the Information Off the Page 76
The Real Difficulties 79
Download the Raw Content If Possible 80
Forms, Dialog Boxes, and New Windows 80
Flash 81
The Dark Side 82
Conclusion 82

6. **Detecting Liars and the Confused in Contradictory Online Reviews................. 83**
Weotta 83
Getting Reviews 84
Sentiment Classification 85
Polarized Language 85
Corpus Creation 87
Training a Classifier 88
Validating the Classifier 90
Designing with Data 91
Lessons Learned 92
Summary 92
Resources 93

7. **Will the Bad Data Please Stand Up?.. 95**
Example 1: Defect Reduction in Manufacturing 95
Example 2: Who's Calling? 98
Example 3: When "Typical" Does Not Mean "Average" 101
Lessons Learned 104
Will This Be on the Test? 105

8. **Blood, Sweat, and Urine.. 107**

A Very Nerdy Body Swap Comedy 107
How Chemists Make Up Numbers 108
All Your Database Are Belong to Us 110
Check, Please 113
Live Fast, Die Young, and Leave a Good-Looking Corpse Code Repository 114
Rehab for Chemists (and Other Spreadsheet Abusers) 115
tl;dr 117

9. When Data and Reality Don't Match.. 119
Whose Ticker Is It Anyway? 120
Splits, Dividends, and Rescaling 122
Bad Reality 125
Conclusion 127

10. Subtle Sources of Bias and Error... 129
Imputation Bias: General Issues 131
Reporting Errors: General Issues 133
Other Sources of Bias 135
  Topcoding/Bottomcoding 136
  Seam Bias 137
  Proxy Reporting 138
  Sample Selection 139
Conclusions 139
References 140

11. Don't Let the Perfect Be the Enemy of the Good: Is Bad Data Really Bad?.......... 143
But First, Let's Reflect on Graduate School ... 143
Moving On to the Professional World 144
Moving into Government Work 146
Government Data Is Very Real 146
Service Call Data as an Applied Example 147
Moving Forward 148
Lessons Learned and Looking Ahead 149

12. When Databases Attack: A Guide for When to Stick to Files....................... 151
History 151
  Building My Toolset 152
  The Roadblock: My Datastore 152
Consider Files as Your Datastore 154
  Files Are Simple! 154
  Files Work with Everything 154
  Files Can Contain Any Data Type 154

Data Corruption Is Local 155
They Have Great Tooling 155
There's No Install Tax 155
File Concepts 156
Encoding 156
Text Files 156
Binary Data 156
Memory-Mapped Files 156
File Formats 156
Delimiters 158
A Web Framework Backed by Files 159
Motivation 160
Implementation 161
Reflections 161

13. Crouching Table, Hidden Network.. 163
A Relational Cost Allocations Model 164
The Delicate Sound of a Combinatorial Explosion... 167
The Hidden Network Emerges 168
Storing the Graph 169
Navigating the Graph with Gremlin 170
Finding Value in Network Properties 171
Think in Terms of Multiple Data Models and Use the Right Tool for the Job 173
Acknowledgments 173

14. Myths of Cloud Computing.. 175
Introduction to the Cloud 175
What Is "The Cloud"? 175
The Cloud and Big Data 176
Introducing Fred 176
At First Everything Is Great 177
They Put 100% of Their Infrastructure in the Cloud 177
As Things Grow, They Scale Easily at First 177
Then Things Start Having Trouble 177
They Need to Improve Performance 178
Higher IO Becomes Critical 178
A Major Regional Outage Causes Massive Downtime 178
Higher IO Comes with a Cost 179
Data Sizes Increase 179
Geo Redundancy Becomes a Priority 179
Horizontal Scale Isn't as Easy as They Hoped 180
Costs Increase Dramatically 180

Fred's Follies 181
Myth 1: Cloud Is a Great Solution for All Infrastructure Components 181
How This Myth Relates to Fred's Story 181
Myth 2: Cloud Will Save Us Money 181
How This Myth Relates to Fred's Story 183
Myth 3: Cloud IO Performance Can Be Improved to Acceptable Levels Through Software RAID 183
How This Myth Relates to Fred's Story 183
Myth 4: Cloud Computing Makes Horizontal Scaling Easy 184
How This Myth Relates to Fred's Story 184
Conclusion and Recommendations 184

15. The Dark Side of Data Science.. 187
Avoid These Pitfalls 187
Know Nothing About Thy Data 188
Be Inconsistent in Cleaning and Organizing the Data 188
Assume Data Is Correct and Complete 188
Spillover of Time-Bound Data 189
Thou Shalt Provide Your Data Scientists with a Single Tool for All Tasks 189
Using a Production Environment for Ad-Hoc Analysis 189
The Ideal Data Science Environment 190
Thou Shalt Analyze for Analysis' Sake Only 191
Thou Shalt Compartmentalize Learnings 192
Thou Shalt Expect Omnipotence from Data Scientists 192
Where Do Data Scientists Live Within the Organization? 193
Final Thoughts 193

16. How to Feed and Care for Your Machine-Learning Experts...................... 195
Define the Problem 195
Fake It Before You Make It 196
Create a Training Set 197
Pick the Features 198
Encode the Data 199
Split Into Training, Test, and Solution Sets 200
Describe the Problem 201
Respond to Questions 201
Integrate the Solutions 202
Conclusion 203

17. Data Traceability... 205
Why? 205
Personal Experience 206

Snapshotting 206
Saving the Source 206
Weighting Sources 207
Backing Out Data 207
Separating Phases (and Keeping them Pure) 207
Identifying the Root Cause 208
Finding Areas for Improvement 208
Immutability: Borrowing an Idea from Functional Programming 208
An Example 209
Crawlers 210
Change 210
Clustering 210
Popularity 210
Conclusion 211

18. Social Media: Erasable Ink? **213**
Social Media: Whose Data Is This Anyway? 214
Control 215
Commercial Resyndication 216
Expectations Around Communication and Expression 217
Technical Implications of New End User Expectations 219
What Does the Industry Do? 221
Validation API 222
Update Notification API 222
What Should End Users Do? 222
How Do We Work Together? 223

19. Data Quality Analysis Demystified: Knowing When Your Data Is Good Enough **225**
Framework Introduction: The Four Cs of Data Quality Analysis 226
Complete 227
Coherent 229
Correct 232
aCcountable 233
Conclusion 237

Index **239**