

Contents

About the Author xix

Author's Preface to Second Edition xxi

Author's Preface to First Edition xxv

1. Introduction	1
Section 1.1. Definition of Big Data	1
Section 1.2. Big Data Versus Small Data	3
Section 1.3. Whence Comest Big Data?	5
Section 1.4. The Most Common Purpose of Big Data Is to Produce Small Data	7
Section 1.5. Big Data Sits at the Center of the Research Universe	8
Glossary	9
References	13
2. Providing Structure to Unstructured Data	15
Section 2.1. Nearly All Data Is Unstructured and Unusable in Its Raw Form	15
Section 2.2. Concordances	16
Section 2.3. Term Extraction	19
Section 2.4. Indexing	22
Section 2.5. Autocoding	24
Section 2.6. Case Study: Instantly Finding the Precise Location of Any Atom in the Universe (Some Assembly Required)	29
Section 2.7. Case Study (Advanced): A Complete Autocoder (in 12 Lines of Python Code)	31
Section 2.8. Case Study: Concordances as Transformations of Text	34

Section 2.9. Case Study (Advanced): Burrows Wheeler Transform (BWT)	36
Glossary	39
References	50
3. Identification, Deidentification, and Reidentification	53
Section 3.1. What Are Identifiers?	53
Section 3.2. Difference Between an Identifier and an Identifier System	55
Section 3.3. Generating Unique Identifiers	58
Section 3.4. Really Bad Identifier Methods	60
Section 3.5. Registering Unique Object Identifiers	63
Section 3.6. Deidentification and Reidentification	66
Section 3.7. Case Study: Data Scrubbing	69
Section 3.8. Case Study (Advanced): Identifiers in Image Headers	71
Section 3.9. Case Study: One-Way Hashes	74
Glossary	76
References	82
4. Metadata, Semantics, and Triples	85
Section 4.1. Metadata	85
Section 4.2. eXtensible Markup Language	85
Section 4.3. Semantics and Triples	87
Section 4.4. Namespaces	88
Section 4.5. Case Study: A Syntax for Triples	90
Section 4.6. Case Study: Dublin Core	93
Glossary	94
References	95

5. Classifications and Ontologies	97
Section 5.1. It's All About Object Relationships	97
Section 5.2. Classifications, the Simplest of Ontologies	101
Section 5.3. Ontologies, Classes With Multiple Parents	104
Section 5.4. Choosing a Class Model	106
Section 5.5. Class Blending	110
Section 5.6. Common Pitfalls in Ontology Development	111
Section 5.7. Case Study: An Upper Level Ontology	114
Section 5.8. Case Study (Advanced): Paradoxes	115
Section 5.9. Case Study (Advanced): RDF Schemas and Class Properties	117
Section 5.10. Case Study (Advanced): Visualizing Class Relationships	120
Glossary	125
References	134
6. Introspection	137
Section 6.1. Knowledge of Self	137
Section 6.2. Data Objects: The Essential Ingredient of Every Big Data Collection	140
Section 6.3. How Big Data Uses Introspection	142
Section 6.4. Case Study: Time Stamping Data	145
Section 6.5. Case Study: A Visit to the TripleStore	147
Section 6.6. Case Study (Advanced): Proof That Big Data Must Be Object-Oriented	152
Glossary	153
References	154
7. Standards and Data Integration	155
Section 7.1. Standards	155
Section 7.2. Specifications Versus Standards	160

Section 7.3. Versioning	162
Section 7.4. Compliance Issues	164
Section 7.5. Case Study: Standardizing the Chocolate Teapot	165
Glossary	166
References	167
8. Immutability and Immortality	169
Section 8.1. The Importance of Data That Cannot Change	169
Section 8.2. Immutability and Identifiers	170
Section 8.3. Coping With the Data That Data Creates	173
Section 8.4. Reconciling Identifiers Across Institutions	174
Section 8.5. Case Study: The Trusted Timestamp	176
Section 8.6. Case Study: Blockchains and Distributed Ledgers	176
Section 8.7. Case Study (Advanced): Zero-Knowledge Reconciliation	179
Glossary	181
References	183
9. Assessing the Adequacy of a Big Data Resource	185
Section 9.1. Looking at the Data	185
Section 9.2. The Minimal Necessary Properties of Big Data	192
Section 9.3. Data That Comes With Conditions	197
Section 9.4. Case Study: Utilities for Viewing and Searching Large Files	198
Section 9.5. Case Study: Flattened Data	200
Glossary	201
References	205
10. Measurement	207
Section 10.1. Accuracy and Precision	207
Section 10.2. Data Range	209

Section 10.3. Counting	211
Section 10.4. Normalizing and Transforming Your Data	215
Section 10.5. Reducing Your Data	219
Section 10.6. Understanding Your Control	222
Section 10.7. Statistical Significance Without Practical Significance	223
Section 10.8. Case Study: Gene Counting	224
Section 10.9. Case Study: Early Biometrics, and the Significance of Narrow Data Ranges	225
Glossary	226
References	228
11. Indispensable Tips for Fast and Simple Big Data Analysis	231
Section 11.1. Speed and Scalability	231
Section 11.2. Fast Operations, Suitable for Big Data, That Every Computer Supports	237
Section 11.3. The Dot Product, a Simple and Fast Correlation Method	243
Section 11.4. Clustering	245
Section 11.5. Methods for Data Persistence (Without Using a Database)	247
Section 11.6. Case Study: Climbing a Classification	249
Section 11.7. Case Study (Advanced): A Database Example	251
Section 11.8. Case Study (Advanced): NoSQL	252
Glossary	253
References	256
12. Finding the Clues in Large Collections of Data	259
Section 12.1. Denominators	259
Section 12.2. Word Frequency Distributions	260
Section 12.3. Outliers and Anomalies	264

Section 12.4. Back-of-Envelope Analyses	266
Section 12.5. Case Study: Predicting User Preferences	268
Section 12.6. Case Study: Multimodality in Population Data	270
Section 12.7. Case Study: Big and Small Black Holes	271
Glossary	271
References	275
13. Using Random Numbers to Knock Your Big Data Analytic Problems Down to Size	277
Section 13.1. The Remarkable Utility of (Pseudo)Random Numbers	277
Section 13.2. Repeated Sampling	283
Section 13.3. Monte Carlo Simulations	288
Section 13.4. Case Study: Proving the Central Limit Theorem	291
Section 13.5. Case Study: Frequency of Unlikely String of Occurrences	293
Section 13.6. Case Study: The Infamous Birthday Problem	294
Section 13.7. Case Study (Advanced): The Monty Hall Problem	295
Section 13.8. Case Study (Advanced): A Bayesian Analysis	297
Glossary	299
References	301
14. Special Considerations in Big Data Analysis	303
Section 14.1. Theory in Search of Data	303
Section 14.2. Data in Search of Theory	304
Section 14.3. Bigness Biases	305
Section 14.4. Data Subsets in Big Data: Neither Additive Nor Transitive	310
Section 14.5. Additional Big Data Pitfalls	311
Section 14.6. Case Study (Advanced): Curse of Dimensionality	314
Glossary	316
References	318

15. Big Data Failures and How to Avoid (Some of) Them	321
Section 15.1. Failure Is Common	321
Section 15.2. Failed Standards	323
Section 15.3. Blaming Complexity	326
Section 15.4. An Approach to Big Data That May Work for You	328
Section 15.5. After Failure	337
Section 15.6. Case Study: Cancer Biomedical Informatics Grid, a Bridge Too Far	339
Section 15.7. Case Study: The Gaussian Copula Function	344
Glossary	345
References	347
16. Data Reanalysis: Much More Important Than Analysis	351
Section 16.1. First Analysis (Nearly) Always Wrong	351
Section 16.2. Why Reanalysis Is More Important Than Analysis	354
Section 16.3. Case Study: Reanalysis of Old JADE Collider Data	356
Section 16.4. Case Study: Vindication Through Reanalysis	357
Section 16.5. Case Study: Finding New Planets From Old Data	357
Glossary	359
References	359
17. Repurposing Big Data	363
Section 17.1. What Is Data Repurposing?	363
Section 17.2. Dark Data, Abandoned Data, and Legacy Data	365
Section 17.3. Case Study: From Postal Code to Demographic Keystone	367
Section 17.4. Case Study: Scientific Inferencing From a Database of Genetic Sequences	368
Section 17.5. Case Study: Linking Global Warming to High-Intensity Hurricanes	369

Section 17.6. Case Study: Inferring Climate Trends With Geologic Data	369
Section 17.7. Case Study: Lunar Orbiter Image Recovery Project	370
Glossary	371
References	372
18. Data Sharing and Data Security	373
Section 18.1. What Is Data Sharing, and Why Don't We Do More of It?	373
Section 18.2. Common Complaints	374
Section 18.3. Data Security and Cryptographic Protocols	381
Section 18.4. Case Study: Life on Mars	387
Section 18.5. Case Study: Personal Identifiers	388
Glossary	390
References	391
19. Legalities	395
Section 19.1. Responsibility for the Accuracy and Legitimacy of Data	395
Section 19.2. Rights to Create, Use, and Share the Resource	398
Section 19.3. Copyright and Patent Infringements Incurred by Using Standards	400
Section 19.4. Protections for Individuals	402
Section 19.5. Consent	404
Section 19.6. Unconsented Data	409
Section 19.7. Privacy Policies	411
Section 19.8. Case Study: Timely Access to Big Data	412
Section 19.9. Case Study: The Havasupai Story	413
Glossary	415
References	416

20. Societal Issues	419
Section 20.1. How Big Data Is Perceived by the Public	419
Section 20.2. Reducing Costs and Increasing Productivity With Big Data	422
Section 20.3. Public Mistrust	424
Section 20.4. Saving Us From Ourselves	425
Section 20.5. Who Is Big Data?	428
Section 20.6. Hubris and Hyperbole	434
Section 20.7. Case Study: The Citizen Scientists	437
Section 20.8. Case Study: 1984, by George Orwell	440
Glossary	441
References	442
Index	445