

Contents

Acknowledgments	xiii
Foreword	xxiii
Introduction	xxvii
<hr/>	
PART ONE	
Thinking Like a Data Head	
CHAPTER 1	
What Is the Problem?	3
Questions a Data Head Should Ask	4
<i>Why Is This Problem Important?</i>	4
<i>Who Does This Problem Affect?</i>	6
<i>What If We Don't Have the Right Data?</i>	6
<i>When Is the Project Over?</i>	7
<i>What If We Don't Like the Results?</i>	7
Understanding Why Data Projects Fail	8
<i>Customer Perception</i>	8
<i>Discussion</i>	10
Working on Problems That Matter	11
Chapter Summary	11
CHAPTER 2	
What Is Data?	13
Data vs. Information	13
<i>An Example Dataset</i>	14
Data Types	15
How Data Is Collected and Structured	16
<i>Observational vs. Experimental Data</i>	16
<i>Structured vs. Unstructured Data</i>	17
Basic Summary Statistics	18
Chapter Summary	19
	xv

CHAPTER 3**Prepare to Think Statistically****21**

Ask Questions 22

There Is Variation in All Things 23

Scenario: Customer Perception (The Sequel) 24*Case Study: Kidney-Cancer Rates* 26

Probabilities and Statistics 28

Probability vs. Intuition 29*Discovery with Statistics* 31

Chapter Summary 33

PART TWO**Speaking Like a Data Head****CHAPTER 4****Argue with the Data****37**

What Would You Do? 38

Missing Data Disaster 39

Tell Me the Data Origin Story 43

Who Collected the Data? 44*How Was the Data Collected?* 44

Is the Data Representative? 45

Is There Sampling Bias? 46*What Did You Do with Outliers?* 46

What Data Am I Not Seeing? 47

How Did You Deal with Missing Values? 47*Can the Data Measure What You Want It to Measure?* 48

Argue with Data of All Sizes 48

Chapter Summary 49

CHAPTER 5**Explore the Data****51**

Exploratory Data Analysis and You 52

Embracing the Exploratory Mindset 52

Questions to Guide You 53*The Setup* 53

Can the Data Answer the Question? 54

Set Expectations and Use Common Sense 54*Do the Values Make Intuitive Sense?* 54*Watch Out: Outliers and Missing Values* 58

92	Did You Discover Any Relationships?	59
93	<i>Understanding Correlation</i>	59
93	<i>Watch Out: Misinterpreting Correlation</i>	60
94	<i>Watch Out: Correlation Does Not Imply Causation</i>	62
95	Did You Find New Opportunities in the Data?	63
96	Chapter Summary	63
CHAPTER 6		
	Examine the Probabilities	65
	Take a Guess	66
	The Rules of the Game	66
	<i>Notation</i>	67
	<i>Conditional Probability and Independent Events</i>	69
	<i>The Probability of Multiple Events</i>	69
	Two Things That Happen Together	69
	One Thing or the Other	70
	Probability Thought Exercise	72
	<i>Next Steps</i>	73
	Be Careful Assuming Independence	74
	<i>Don't Fall for the Gambler's Fallacy</i>	74
	All Probabilities Are Conditional	75
	<i>Don't Swap Dependencies</i>	76
	<i>Bayes' Theorem</i>	76
	Ensure the Probabilities Have Meaning	79
	<i>Calibration</i>	80
	<i>Rare Events Can, and Do, Happen</i>	80
	Chapter Summary	81
CHAPTER 7		
	Challenge the Statistics	83
	Quick Lessons on Inference	83
	<i>Give Yourself Some Wiggle Room</i>	84
	<i>More Data, More Evidence</i>	84
	<i>Challenge the Status Quo</i>	85
	<i>Evidence to the Contrary</i>	86
	<i>Balance Decision Errors</i>	88
	The Process of Statistical Inference	89
	The Questions You Should Ask to Challenge the Statistics	90
	<i>What Is the Context for These Statistics?</i>	90
	<i>What Is the Sample Size?</i>	91
	<i>What Are You Testing?</i>	92

<i>What Is the Null Hypothesis?</i>	92
Assuming Equivalence	93
<i>What Is the Significance Level?</i>	93
<i>How Many Tests Are You Doing?</i>	94
<i>Can I See the Confidence Intervals?</i>	95
<i>Is This Practically Significant?</i>	96
<i>Are You Assuming Causality?</i>	96
Chapter Summary	97

PART THREE

Understanding the Data Scientist's Toolbox

CHAPTER 8

Search for Hidden Groups	101
Unsupervised Learning	102
Dimensionality Reduction	102
<i>Creating Composite Features</i>	103
Principal Component Analysis	105
<i>Principal Components in Athletic Ability</i>	105
<i>PCA Summary</i>	108
<i>Potential Traps</i>	109
Clustering	110
<i>k</i> -Means Clustering	111
<i>Clustering Retail Locations</i>	111
<i>Potential Traps</i>	113
Chapter Summary	114

CHAPTER 9

Understand the Regression Model	117
Supervised Learning	117
Linear Regression: What It Does	119
<i>Least Squares Regression: Not Just a Clever Name</i>	120
Linear Regression: What It Gives You	123
<i>Extending to Many Features</i>	124
Linear Regression: What Confusion It Causes	125
<i>Omitted Variables</i>	125
<i>Multicollinearity</i>	126
<i>Data Leakage</i>	127
<i>Extrapolation Failures</i>	128

<i>Many Relationships Aren't Linear</i>	128
<i>Are You Explaining or Predicting?</i>	128
<i>Regression Performance</i>	130
Other Regression Models	131
Chapter Summary	131
CHAPTER 10	
Understand the Classification Model	133
Introduction to Classification	133
<i>What You'll Learn</i>	134
<i>Classification Problem Setup</i>	135
Logistic Regression	135
<i>Logistic Regression: So What?</i>	138
Decision Trees	139
Ensemble Methods	142
<i>Random Forests</i>	143
<i>Gradient Boosted Trees</i>	143
<i>Interpretability of Ensemble Models</i>	145
Watch Out for Pitfalls	145
<i>Misapplication of the Problem</i>	146
<i>Data Leakage</i>	146
<i>Not Splitting Your Data</i>	146
<i>Choosing the Right Decision Threshold</i>	147
Misunderstanding Accuracy	147
<i>Confusion Matrices</i>	148
Chapter Summary	150
CHAPTER 11	
Understand Text Analytics	151
Expectations of Text Analytics	151
How Text Becomes Numbers	153
<i>A Big Bag of Words</i>	153
<i>N-Grams</i>	157
<i>Word Embeddings</i>	158
Topic Modeling	160
Text Classification	163
<i>Naïve Bayes</i>	164
<i>Sentiment Analysis</i>	166
Practical Considerations When Working with Text	167
<i>Big Tech Has the Upper Hand</i>	168
Chapter Summary	169

CHAPTER 12		
 Conceptualize Deep Learning		171
Neural Networks		172
<i>How Are Neural Networks Like the Brain?</i>		172
<i>A Simple Neural Network</i>		173
<i>How a Neural Network Learns</i>		174
<i>A Slightly More Complex Neural Network</i>		175
Applications of Deep Learning		178
<i>The Benefits of Deep Learning</i>		179
<i>How Computers “See” Images</i>		180
<i>Convolutional Neural Networks</i>		182
<i>Deep Learning on Language and Sequences</i>		183
Deep Learning in Practice		185
<i>Do You Have Data?</i>		185
<i>Is Your Data Structured?</i>		186
<i>What Will the Network Look Like?</i>		186
Artificial Intelligence and You		187
<i>Big Tech Has the Upper Hand</i>		188
<i>Ethics in Deep Learning</i>		189
Chapter Summary		190
PART FOUR		
 Ensuring Success		
CHAPTER 13		
 Watch Out for Pitfalls		193
Biases and Weird Phenomena in Data		194
<i>Survivorship Bias</i>		194
<i>Regression to the Mean</i>		195
<i>Simpson’s Paradox</i>		195
<i>Confirmation Bias</i>		197
<i>Effort Bias (aka the “Sunk Cost Fallacy”)</i>		197
<i>Algorithmic Bias</i>		198
<i>Uncategorized Bias</i>		198
The Big List of Pitfalls		199
<i>Statistical and Machine Learning Pitfalls</i>		199
<i>Project Pitfalls</i>		200
Chapter Summary		202

CHAPTER 14	
Know the People and Personalities	203
Seven Scenes of Communication Breakdowns	204
<i>The Postmortem</i>	204
<i>Storytime</i>	205
<i>The Telephone Game</i>	206
<i>Into the Weeds</i>	206
<i>The Reality Check</i>	207
<i>The Takeover</i>	207
<i>The Blowhard</i>	208
Data Personalities	208
<i>Data Enthusiasts</i>	209
<i>Data Cynics</i>	209
<i>Data Heads</i>	209
Chapter Summary	210
CHAPTER 15	
What's Next?	211
Index	215