

# Table of Contents

<b>Preface</b> .....	<b>xi</b>
<b>1. Introduction: Data-Analytic Thinking</b> .....	<b>1</b>
The Ubiquity of Data Opportunities	1
Example: Hurricane Frances	3
Example: Predicting Customer Churn	4
Data Science, Engineering, and Data-Driven Decision Making	4
Data Processing and “Big Data”	7
From Big Data 1.0 to Big Data 2.0	8
Data and Data Science Capability as a Strategic Asset	9
Data-Analytic Thinking	12
This Book	14
Data Mining and Data Science, Revisited	14
Chemistry Is Not About Test Tubes: Data Science Versus the Work of the Data Scientist	15
Summary	16
<b>2. Business Problems and Data Science Solutions</b> .....	<b>19</b>
<i>Fundamental concepts: A set of canonical data mining tasks; The data mining process;     Supervised versus unsupervised data mining.</i>	
From Business Problems to Data Mining Tasks	19
Supervised Versus Unsupervised Methods	24
Data Mining and Its Results	25
The Data Mining Process	26
Business Understanding	28
Data Understanding	28
Data Preparation	30
Modeling	31
Evaluation	31

Deployment	32
Implications for Managing the Data Science Team	34
Other Analytics Techniques and Technologies	35
Statistics	35
Database Querying	37
Data Warehousing	38
Regression Analysis	39
Machine Learning and Data Mining	39
Answering Business Questions with These Techniques	40
Summary	41
<b>3. Introduction to Predictive Modeling: From Correlation to Supervised Segmentation.</b>	<b>43</b>
<i>Fundamental concepts: Identifying informative attributes; Segmenting data by progressive attribute selection.</i>	
<i>Exemplary techniques: Finding correlations; Attribute/variable selection; Tree induction.</i>	
Models, Induction, and Prediction	44
Supervised Segmentation	48
Selecting Informative Attributes	49
Example: Attribute Selection with Information Gain	56
Supervised Segmentation with Tree-Structured Models	62
Visualizing Segmentations	67
Trees as Sets of Rules	71
Probability Estimation	71
Example: Addressing the Churn Problem with Tree Induction	73
Summary	78
<b>4. Fitting a Model to Data.</b>	<b>81</b>
<i>Fundamental concepts: Finding "optimal" model parameters based on data; Choosing the goal for data mining; Objective functions; Loss functions.</i>	
<i>Exemplary techniques: Linear regression; Logistic regression; Support-vector machines.</i>	
Classification via Mathematical Functions	83
Linear Discriminant Functions	85
Optimizing an Objective Function	87
An Example of Mining a Linear Discriminant from Data	88
Linear Discriminant Functions for Scoring and Ranking Instances	90
Support Vector Machines, Briefly	91
Regression via Mathematical Functions	94
Class Probability Estimation and Logistic "Regression"	96
* Logistic Regression: Some Technical Details	99
Example: Logistic Regression versus Tree Induction	102
Nonlinear Functions, Support Vector Machines, and Neural Networks	105

<b>5. Overfitting and Its Avoidance.....</b>	<b>111</b>
<i>Fundamental concepts: Generalization; Fitting and overfitting; Complexity control.</i>	
<i>Exemplary techniques: Cross-validation; Attribute selection; Tree pruning; Regularization.</i>	
Generalization	111
Overfitting	113
Overfitting Examined	113
Holdout Data and Fitting Graphs	113
Overfitting in Tree Induction	116
Overfitting in Mathematical Functions	118
Example: Overfitting Linear Functions	119
* Example: Why Is Overfitting Bad?	124
From Holdout Evaluation to Cross-Validation	126
The Churn Dataset Revisited	129
Learning Curves	130
Overfitting Avoidance and Complexity Control	133
Avoiding Overfitting with Tree Induction	133
A General Method for Avoiding Overfitting	134
* Avoiding Overfitting for Parameter Optimization	136
Summary	140
<b>6. Similarity, Neighbors, and Clusters.....</b>	<b>141</b>
<i>Fundamental concepts: Calculating similarity of objects described by data; Using similarity for prediction; Clustering as similarity-based segmentation.</i>	
<i>Exemplary techniques: Searching for similar entities; Nearest neighbor methods; Clustering methods; Distance metrics for calculating similarity.</i>	
Similarity and Distance	142
Nearest-Neighbor Reasoning	144
Example: Whiskey Analytics	144
Nearest Neighbors for Predictive Modeling	146
How Many Neighbors and How Much Influence?	149
Geometric Interpretation, Overfitting, and Complexity Control	151
Issues with Nearest-Neighbor Methods	154
Some Important Technical Details Relating to Similarities and Neighbors	157
Heterogeneous Attributes	157
* Other Distance Functions	158
* Combining Functions: Calculating Scores from Neighbors	161
Clustering	163
Example: Whiskey Analytics Revisited	163
Hierarchical Clustering	164

Nearest Neighbors Revisited: Clustering Around Centroids	169
Example: Clustering Business News Stories	174
Understanding the Results of Clustering	177
* Using Supervised Learning to Generate Cluster Descriptions	179
Stepping Back: Solving a Business Problem Versus Data Exploration	182
Summary	184
<b>7. Decision Analytic Thinking I: What Is a Good Model?.....</b>	<b>187</b>
<i>Fundamental concepts: Careful consideration of what is desired from data science results; Expected value as a key evaluation framework; Consideration of appropriate comparative baselines.</i>	
<i>Exemplary techniques: Various evaluation metrics; Estimating costs and benefits; Calculating expected profit; Creating baseline methods for comparison.</i>	
Evaluating Classifiers	188
Plain Accuracy and Its Problems	189
The Confusion Matrix	189
Problems with Unbalanced Classes	190
Problems with Unequal Costs and Benefits	193
Generalizing Beyond Classification	193
A Key Analytical Framework: Expected Value	194
Using Expected Value to Frame Classifier Use	195
Using Expected Value to Frame Classifier Evaluation	196
Evaluation, Baseline Performance, and Implications for Investments in Data	204
Summary	207
<b>8. Visualizing Model Performance.....</b>	<b>209</b>
<i>Fundamental concepts: Visualization of model performance under various kinds of uncertainty; Further consideration of what is desired from data mining results.</i>	
<i>Exemplary techniques: Profit curves; Cumulative response curves; Lift curves; ROC curves.</i>	
Ranking Instead of Classifying	209
Profit Curves	212
ROC Graphs and Curves	214
The Area Under the ROC Curve (AUC)	219
Cumulative Response and Lift Curves	219
Example: Performance Analytics for Churn Modeling	223
Summary	231
<b>9. Evidence and Probabilities.....</b>	<b>233</b>
<i>Fundamental concepts: Explicit evidence combination with Bayes' Rule; Probabilistic reasoning via assumptions of conditional independence.</i>	
<i>Exemplary techniques: Naive Bayes classification; Evidence lift.</i>	

Example: Targeting Online Consumers With Advertisements	233
Combining Evidence Probabilistically	235
Joint Probability and Independence	236
Bayes' Rule	237
Applying Bayes' Rule to Data Science	239
Conditional Independence and Naive Bayes	240
Advantages and Disadvantages of Naive Bayes	242
A Model of Evidence "Lift"	244
Example: Evidence Lifts from Facebook "Likes"	245
Evidence in Action: Targeting Consumers with Ads	247
Summary	247
<b>10. Representing and Mining Text. ....</b>	<b>249</b>
<i>Fundamental concepts: The importance of constructing mining-friendly data representations; Representation of text for data mining.</i>	
<i>Exemplary techniques: Bag of words representation; TFIDF calculation; N-grams; Stemming; Named entity extraction; Topic models.</i>	
Why Text Is Important	250
Why Text Is Difficult	250
Representation	251
Bag of Words	252
Term Frequency	252
Measuring Sparseness: Inverse Document Frequency	254
Combining Them: TFIDF	256
Example: Jazz Musicians	256
* The Relationship of IDF to Entropy	261
Beyond Bag of Words	263
N-gram Sequences	263
Named Entity Extraction	264
Topic Models	264
Example: Mining News Stories to Predict Stock Price Movement	266
The Task	266
The Data	268
Data Preprocessing	271
Results	271
Summary	275
<b>11. Decision Analytic Thinking II: Toward Analytical Engineering. ....</b>	<b>277</b>
<i>Fundamental concept: Solving business problems with data science starts with analytical engineering: designing an analytical solution, based on the data, tools, and techniques available.</i>	
<i>Exemplary technique: Expected value as a framework for data science solution design.</i>	

Targeting the Best Prospects for a Charity Mailing	278
The Expected Value Framework: Decomposing the Business Problem and Recomposing the Solution Pieces	278
A Brief Digression on Selection Bias	280
Our Churn Example Revisited with Even More Sophistication	281
The Expected Value Framework: Structuring a More Complicated Business Problem	281
Assessing the Influence of the Incentive	283
From an Expected Value Decomposition to a Data Science Solution	284
Summary	287
<b>12. Other Data Science Tasks and Techniques.....</b>	<b>289</b>
<i>Fundamental concepts: Our fundamental concepts as the basis of many common data science techniques; The importance of familiarity with the building blocks of data science.</i>	
<i>Exemplary techniques: Association and co-occurrences; Behavior profiling; Link prediction; Data reduction; Latent information mining; Movie recommendation; Bias-variance decomposition of error; Ensembles of models; Causal reasoning from data.</i>	
Co-occurrences and Associations: Finding Items That Go Together	290
Measuring Surprise: Lift and Leverage	291
Example: Beer and Lottery Tickets	292
Associations Among Facebook Likes	293
Profiling: Finding Typical Behavior	296
Link Prediction and Social Recommendation	301
Data Reduction, Latent Information, and Movie Recommendation	302
Bias, Variance, and Ensemble Methods	306
Data-Driven Causal Explanation and a Viral Marketing Example	309
Summary	310
<b>13. Data Science and Business Strategy.....</b>	<b>313</b>
<i>Fundamental concepts: Our principles as the basis of success for a data-driven business; Acquiring and sustaining competitive advantage via data science; The importance of careful curation of data science capability.</i>	
Thinking Data-Analytically, Redux	313
Achieving Competitive Advantage with Data Science	315
Sustaining Competitive Advantage with Data Science	316
Formidable Historical Advantage	317
Unique Intellectual Property	317
Unique Intangible Collateral Assets	318
Superior Data Scientists	318
Superior Data Science Management	320
Attracting and Nurturing Data Scientists and Their Teams	321

Examine Data Science Case Studies	323
Be Ready to Accept Creative Ideas from Any Source	324
Be Ready to Evaluate Proposals for Data Science Projects	324
Example Data Mining Proposal	325
Flaws in the Big Red Proposal	326
A Firm's Data Science Maturity	327
<b>14. Conclusion.....</b>	<b>331</b>
The Fundamental Concepts of Data Science	331
Applying Our Fundamental Concepts to a New Problem: Mining Mobile	
Device Data	334
Changing the Way We Think about Solutions to Business Problems	337
What Data Can't Do: Humans in the Loop, Revisited	338
Privacy, Ethics, and Mining Data About Individuals	341
Is There More to Data Science?	342
Final Example: From Crowd-Sourcing to Cloud-Sourcing	343
Final Words	344
<b>A. Proposal Review Guide.....</b>	<b>347</b>
<b>B. Another Sample Proposal.....</b>	<b>351</b>
<b>Glossary.....</b>	<b>355</b>
<b>Bibliography.....</b>	<b>359</b>
<b>Index.....</b>	<b>367</b>