

Contents

1	What is Data Science?	1
1.1	Computer Science, Data Science, and Real Science	2
1.2	Asking Interesting Questions from Data	4
1.2.1	The Baseball Encyclopedia	5
1.2.2	The Internet Movie Database (IMDb)	7
1.2.3	Google Ngrams	10
1.2.4	New York Taxi Records	11
1.3	Properties of Data	14
1.3.1	Structured vs. Unstructured Data	14
1.3.2	Quantitative vs. Categorical Data	15
1.3.3	Big Data vs. Little Data	15
1.4	Classification and Regression	16
1.5	Data Science Television: The Quant Shop	17
1.5.1	Kaggle Challenges	19
1.6	About the War Stories	19
1.7	War Story: Answering the Right Question	21
1.8	Chapter Notes	22
1.9	Exercises	23
2	Mathematical Preliminaries	27
2.1	Probability	27
2.1.1	Probability vs. Statistics	29
2.1.2	Compound Events and Independence	30
2.1.3	Conditional Probability	31
2.1.4	Probability Distributions	32
2.2	Descriptive Statistics	34
2.2.1	Centrality Measures	34
2.2.2	Variability Measures	36
2.2.3	Interpreting Variance	37
2.2.4	Characterizing Distributions	39
2.3	Correlation Analysis	40
2.3.1	Correlation Coefficients: Pearson and Spearman Rank . .	41
2.3.2	The Power and Significance of Correlation	43
2.3.3	Correlation Does Not Imply Causation!	45

2.3.4	Detecting Periodicities by Autocorrelation	46
2.4	Logarithms	47
2.4.1	Logarithms and Multiplying Probabilities	48
2.4.2	Logarithms and Ratios	48
2.4.3	Logarithms and Normalizing Skewed Distributions	49
2.5	War Story: Fitting Designer Genes	50
2.6	Chapter Notes	52
2.7	Exercises	53
3	Data Munging	57
3.1	Languages for Data Science	57
3.1.1	The Importance of Notebook Environments	59
3.1.2	Standard Data Formats	61
3.2	Collecting Data	64
3.2.1	Hunting	64
3.2.2	Scraping	67
3.2.3	Logging	68
3.3	Cleaning Data	69
3.3.1	Errors vs. Artifacts	69
3.3.2	Data Compatibility	72
3.3.3	Dealing with Missing Values	76
3.3.4	Outlier Detection	78
3.4	War Story: Beating the Market	79
3.5	Crowdsourcing	80
3.5.1	The Penny Demo	81
3.5.2	When is the Crowd Wise?	82
3.5.3	Mechanisms for Aggregation	83
3.5.4	Crowdsourcing Services	84
3.5.5	Gamification	88
3.6	Chapter Notes	90
3.7	Exercises	90
4	Scores and Rankings	95
4.1	The Body Mass Index (BMI)	96
4.2	Developing Scoring Systems	99
4.2.1	Gold Standards and Proxies	99
4.2.2	Scores vs. Rankings	100
4.2.3	Recognizing Good Scoring Functions	101
4.3	Z-scores and Normalization	103
4.4	Advanced Ranking Techniques	104
4.4.1	Elo Rankings	104
4.4.2	Merging Rankings	108
4.4.3	Digraph-based Rankings	109
4.4.4	PageRank	111
4.5	War Story: Clyde's Revenge	111
4.6	Arrow's Impossibility Theorem	114

4.7	War Story: Who's Bigger?	115
4.8	Chapter Notes	118
4.9	Exercises	119
5	Statistical Analysis	121
5.1	Statistical Distributions	122
5.1.1	The Binomial Distribution	123
5.1.2	The Normal Distribution	124
5.1.3	Implications of the Normal Distribution	126
5.1.4	Poisson Distribution	127
5.1.5	Power Law Distributions	129
5.2	Sampling from Distributions	132
5.2.1	Random Sampling beyond One Dimension	133
5.3	Statistical Significance	135
5.3.1	The Significance of Significance	135
5.3.2	The T-test: Comparing Population Means	137
5.3.3	The Kolmogorov-Smirnov Test	139
5.3.4	The Bonferroni Correction	141
5.3.5	False Discovery Rate	142
5.4	War Story: Discovering the Fountain of Youth?	143
5.5	Permutation Tests and P-values	145
5.5.1	Generating Random Permutations	147
5.5.2	DiMaggio's Hitting Streak	148
5.6	Bayesian Reasoning	150
5.7	Chapter Notes	151
5.8	Exercises	151
6	Visualizing Data	155
6.1	Exploratory Data Analysis	156
6.1.1	Confronting a New Data Set	156
6.1.2	Summary Statistics and Anscombe's Quartet	159
6.1.3	Visualization Tools	160
6.2	Developing a Visualization Aesthetic	162
6.2.1	Maximizing Data-Ink Ratio	163
6.2.2	Minimizing the Lie Factor	164
6.2.3	Minimizing Chartjunk	165
6.2.4	Proper Scaling and Labeling	167
6.2.5	Effective Use of Color and Shading	168
6.2.6	The Power of Repetition	169
6.3	Chart Types	170
6.3.1	Tabular Data	170
6.3.2	Dot and Line Plots	174
6.3.3	Scatter Plots	177
6.3.4	Bar Plots and Pie Charts	179
6.3.5	Histograms	183
6.3.6	Data Maps	187

6.4	Great Visualizations	189
6.4.1	Marey's Train Schedule	189
6.4.2	Snow's Cholera Map	191
6.4.3	New York's Weather Year	192
6.5	Reading Graphs	192
6.5.1	The Obscured Distribution	193
6.5.2	Overinterpreting Variance	193
6.6	Interactive Visualization	195
6.7	War Story: TextMapping the World	196
6.8	Chapter Notes	198
6.9	Exercises	199
7	Mathematical Models	201
7.1	Philosophies of Modeling	201
7.1.1	Occam's Razor	201
7.1.2	Bias-Variance Trade-Offs	202
7.1.3	What Would Nate Silver Do?	203
7.2	A Taxonomy of Models	205
7.2.1	Linear vs. Non-Linear Models	206
7.2.2	Blackbox vs. Descriptive Models	206
7.2.3	First-Principle vs. Data-Driven Models	207
7.2.4	Stochastic vs. Deterministic Models	208
7.2.5	Flat vs. Hierarchical Models	209
7.3	Baseline Models	210
7.3.1	Baseline Models for Classification	210
7.3.2	Baseline Models for Value Prediction	212
7.4	Evaluating Models	212
7.4.1	Evaluating Classifiers	213
7.4.2	Receiver-Operator Characteristic (ROC) Curves	218
7.4.3	Evaluating Multiclass Systems	219
7.4.4	Evaluating Value Prediction Models	221
7.5	Evaluation Environments	224
7.5.1	Data Hygiene for Evaluation	225
7.5.2	Amplifying Small Evaluation Sets	226
7.6	War Story: 100% Accuracy	228
7.7	Simulation Models	229
7.8	War Story: Calculated Bets	230
7.9	Chapter Notes	233
7.10	Exercises	234
8	Linear Algebra	237
8.1	The Power of Linear Algebra	237
8.1.1	Interpreting Linear Algebraic Formulae	238
8.1.2	Geometry and Vectors	240
8.2	Visualizing Matrix Operations	241
8.2.1	Matrix Addition	242

8.2.2	Matrix Multiplication	243
8.2.3	Applications of Matrix Multiplication	244
8.2.4	Identity Matrices and Inversion	248
8.2.5	Matrix Inversion and Linear Systems	250
8.2.6	Matrix Rank	251
8.3	Factoring Matrices	252
8.3.1	Why Factor Feature Matrices?	252
8.3.2	LU Decomposition and Determinants	254
8.4	Eigenvalues and Eigenvectors	255
8.4.1	Properties of Eigenvalues	255
8.4.2	Computing Eigenvalues	256
8.5	Eigenvalue Decomposition	257
8.5.1	Singular Value Decomposition	258
8.5.2	Principal Components Analysis	260
8.6	War Story: The Human Factors	262
8.7	Chapter Notes	263
8.8	Exercises	263
9	Linear and Logistic Regression	267
9.1	Linear Regression	268
9.1.1	Linear Regression and Duality	268
9.1.2	Error in Linear Regression	269
9.1.3	Finding the Optimal Fit	270
9.2	Better Regression Models	272
9.2.1	Removing Outliers	272
9.2.2	Fitting Non-Linear Functions	273
9.2.3	Feature and Target Scaling	274
9.2.4	Dealing with Highly-Correlated Features	277
9.3	War Story: Taxi Driver	277
9.4	Regression as Parameter Fitting	279
9.4.1	Convex Parameter Spaces	280
9.4.2	Gradient Descent Search	281
9.4.3	What is the Right Learning Rate?	283
9.4.4	Stochastic Gradient Descent	285
9.5	Simplifying Models through Regularization	286
9.5.1	Ridge Regression	286
9.5.2	LASSO Regression	287
9.5.3	Trade-Offs between Fit and Complexity	288
9.6	Classification and Logistic Regression	289
9.6.1	Regression for Classification	290
9.6.2	Decision Boundaries	291
9.6.3	Logistic Regression	292
9.7	Issues in Logistic Classification	295
9.7.1	Balanced Training Classes	295
9.7.2	Multi-Class Classification	297
9.7.3	Hierarchical Classification	298

9.7.4	Partition Functions and Multinomial Regression	299
9.8	Chapter Notes	300
9.9	Exercises	301
10	Distance and Network Methods	303
10.1	Measuring Distances	303
10.1.1	Distance Metrics	304
10.1.2	The L_k Distance Metric	305
10.1.3	Working in Higher Dimensions	307
10.1.4	Dimensional Egalitarianism	308
10.1.5	Points vs. Vectors	309
10.1.6	Distances between Probability Distributions	310
10.2	Nearest Neighbor Classification	311
10.2.1	Seeking Good Analogies	312
10.2.2	k -Nearest Neighbors	313
10.2.3	Finding Nearest Neighbors	315
10.2.4	Locality Sensitive Hashing	317
10.3	Graphs, Networks, and Distances	319
10.3.1	Weighted Graphs and Induced Networks	320
10.3.2	Talking About Graphs	321
10.3.3	Graph Theory	323
10.4	PageRank	325
10.5	Clustering	327
10.5.1	k -means Clustering	330
10.5.2	Agglomerative Clustering	336
10.5.3	Comparing Clusterings	341
10.5.4	Similarity Graphs and Cut-Based Clustering	341
10.6	War Story: Cluster Bombing	344
10.7	Chapter Notes	345
10.8	Exercises	346
11	Machine Learning	351
11.1	Naive Bayes	354
11.1.1	Formulation	354
11.1.2	Dealing with Zero Counts (Discounting)	356
11.2	Decision Tree Classifiers	357
11.2.1	Constructing Decision Trees	359
11.2.2	Realizing Exclusive Or	361
11.2.3	Ensembles of Decision Trees	362
11.3	Boosting and Ensemble Learning	363
11.3.1	Voting with Classifiers	363
11.3.2	Boosting Algorithms	364
11.4	Support Vector Machines	366
11.4.1	Linear SVMs	369
11.4.2	Non-linear SVMs	369
11.4.3	Kernels	371

11.5	Degrees of Supervision	372
11.5.1	Supervised Learning	372
11.5.2	Unsupervised Learning	372
11.5.3	Semi-supervised Learning	374
11.5.4	Feature Engineering	375
11.6	Deep Learning	377
11.6.1	Networks and Depth	378
11.6.2	Backpropagation	382
11.6.3	Word and Graph Embeddings	383
11.7	War Story: The Name Game	385
11.8	Chapter Notes	387
11.9	Exercises	388
12	Big Data: Achieving Scale	391
12.1	What is Big Data?	392
12.1.1	Big Data as Bad Data	392
12.1.2	The Three Vs	394
12.2	War Story: Infrastructure Matters	395
12.3	Algorithmics for Big Data	397
12.3.1	Big Oh Analysis	397
12.3.2	Hashing	399
12.3.3	Exploiting the Storage Hierarchy	401
12.3.4	Streaming and Single-Pass Algorithms	402
12.4	Filtering and Sampling	403
12.4.1	Deterministic Sampling Algorithms	404
12.4.2	Randomized and Stream Sampling	406
12.5	Parallelism	406
12.5.1	One, Two, Many	407
12.5.2	Data Parallelism	409
12.5.3	Grid Search	409
12.5.4	Cloud Computing Services	410
12.6	MapReduce	410
12.6.1	Map-Reduce Programming	412
12.6.2	MapReduce under the Hood	414
12.7	Societal and Ethical Implications	416
12.8	Chapter Notes	419
12.9	Exercises	419
13	Coda	423
13.1	Get a Job!	423
13.2	Go to Graduate School!	424
13.3	Professional Consulting Services	425
14	Bibliography	427



<http://www.springer.com/978-3-319-55443-3>

The Data Science Design Manual

Skiena, S.S.

2017, XVII, 445 p. 180 illus., 137 illus. in color.,

Hardcover

ISBN: 978-3-319-55443-3